Using Trees to Capture Reticulate Evolution

Lateral Gene Transfers and Cancer Progression

ALI TOFIGH

Doctoral Thesis
Stockholm, Sweden, 2009

TRITA-CSC-A 2009:10
ISSN-1653-5723
ISRN–KTH/CSC/A–09/10-SE
ISBN 978-91-7415-349-1

KTH School of Computer Science
and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i datalogi fredagen den 12 juni 2009 klockan 10.00 i Svedbergssalen, Albanova, Roslagstullsbacken 21, Stockholm.

Tryck: Universitetsservice US AB

To Bita

Abstract

The historic relationship of species and genes are traditionally depicted using trees. However, not all evolutionary histories are adequately captured by bifurcating processes and an increasing amount of research is devoted towards using networks or network-like structures to capture evolutionary history. Lateral gene transfer (LGT) is a previously controversial mechanism responsible for non tree-like evolutionary histories, and is today accepted as a major force of evolution, particularly in the prokaryotic domain.

In this thesis, we present models of gene evolution incorporating both LGTs and duplications, together with efficient computational methods for various inference problems. Specifically, we define a biologically sound combinatorial model for reconciliation of species and gene trees that facilitates simultaneous consideration of duplications and LGTs. We prove that finding most parsimonious reconciliations is NP-hard, but that the problem can be solved efficiently if reconciliations are not required to be acyclic—a condition that is satisfied when analyzing most real-world datasets. We also provide a polynomial-time algorithm for parametric tree reconciliation, a problem analogous to parametric sequence alignment, that enables us to study the entire space of optimal reconciliations under all possible cost schemes.

Going beyond combinatorial models, we define the first probabilistic model of gene evolution incorporating a birth-death process generating duplications, LGTs, and losses, together with a relaxed molecular clock model of sequence evolution. Algorithms based on Markov chain Monte Carlo (MCMC) techniques, methods from numerical analysis, and dynamic programming are presented for various probability and parameter inference problems.

Finally, we develop methods for analysis of cancer progression, a biological process with many similarities to the process of evolution. Cancer progresses by accumulation of harmful genetic aberrations whose patterns of emergence are graph-like. We develop a model of cancer progression based on trees, and mixtures thereof, that admits an efficient structural EM algorithm for finding Maximum Likelihood (ML) solutions from available cross-sectional data.

Contents

List of Publications

- **Simultaneous Identification of Duplications and Lateral Gene Transfers**
  A. Tofigh, M. Hallett, and J. Lagergren
  Based on [67], which was presented at RECOMB 2004
  submitted

- **Inferring Duplications and Lateral Gene Transfers—An Algorithm for Parametric Tree Reconciliation**
  A. Tofigh and J. Lagergren
  Manuscript

- **Detecting LGTs Using a Novel Probabilistic Model Integrating Duplication, LGTs, Losses, Rate Variation, and Sequence Evolution**
  A. Tofigh, J. Sjöstrand, B. Sennblad, L. Arvestad, and J. Lagergren
  A. Tofigh and J. Sjöstrand have contributed equally to this manuscript
  Manuscript

- **A Global Structural EM algorithm for a Model of Cancer Progression**
  A. Tofigh and J. Lagergren
  Manuscript

Acknowledgments

The journey has been long and laborious, but nonetheless quite memorable. I am grateful to all those who provided support and made the tough times less burdensome, and special thanks go to the people who contributed, both directly and indirectly, to the work presented in this thesis.

Foremost among the latter is my supervisor **Jens Lagergren**. He is truly one of the coolest professors I have met and a true scientist—*never* stressed about research, *always* stressed about administration and teaching duties. He is also one of the fastest thinkers I know. It is only through years of training provided by myself and other students that he has learned to come down to the level of us mere mortals when explaining his scientific ideas. I'm especially grateful for his constant support and calm attitude in the face of impending disasters!

Others who have made contributions to this thesis include **Joel**, who has worked hard during weekends to finish experiments and graphs, and **Lasse** and **Bengt** with whom I've had many scientific discussions.

During the years as a PhD student I have shared office space with some very interesting personalities. **Örjan**, my brother in arms, who managed to write his thesis *and* become a first-time father in a matter of just a few months. A true Swedish hero! **Johannes**, is there *anyone* you don't know or haven't met? **Isaac**, it takes sharp minds and brainy people like yourself to build a search engine whose name is now an official verb. **Marcus**, we never shared office space, but it feels like we did. Finally, there was **Samuel**, who left Stockholm and never looked back.

I'm also glad that I got to meet Jens's new PhD students, **Joel** and **Hossein**. Somehow, the blend of Hossein's evilness, Joel's enthusaism, and my lust for revenge led to one of the most successful practical jokes at our department (sorry Jens, but your reaction was priceless!).

A source of welcome distraction has been the (ir)regular poker nights, a tradition I hope will continue for a long time to come. In short, it has been a pleasure losing to you all: **Håkan, Marcus, Katarina, Per, Pär, Diana, Anna, Andreas, Kristoffer, Jenny, Tomas, Björn, Aron, Maria, Erik Sj**.

Finally, there is the one person without whom I would not have survived through it all, and who has had to put up with my shifting moods, especially towards the end: my lovely wife **Bita**, you are too good to be true!

There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

— Charles Darwin,
On the Origin of Species

Chapter 1

Introduction

Darwin's observations during his five-year voyage on HMS Beagle laid the foundation for the body of work that he would later produce, providing compelling evidence for a process of descent with modification and natural selection. Although evolutionary ideas had been formulated in various forms before Darwin, his work popularized the idea among scientists as well as the general public. Of course, Darwin could not have known that long strands of DNA molecules, and the genes located on them, are the vehicles on which traits are inherited from generation to generation, or that discrete events such as recombination, mutation, duplication, and gene transfer are responsible for modification.

Much of Darwin's observations were based on morphological similarities and dissimilarities between species. In fact, not long ago, classification of species and inference of their evolutionary histories were mainly based on morphological data. With the emergence of modern molecular biology, we have at our disposal a detailed model of the mechanisms of inheritance, and so the study of evolution has largely shifted towards the study of DNA and genes. Methods for the construction of gene trees showing the relationship among homologous genes have been particularly useful. Trees or networks showing the evolutionary history of species are often inferred from sets of gene trees. This problem would be trivial if the history of genes simply followed that of species (or vice versa!), but genome content, and with it, the history of genes, is constantly changing in ways that do not always reflect the history of species. Prominent among these events is gene duplication, which has been extensively studied both biologically and computationally. Another event, one that has risen to prominence more recently, is lateral gene transfer (LGT), also known as horizontal gene transfer (HGT). Many computational methods have been developed that incorporate either duplications or LGTs, but few have attempted to incorporate both. A major part of this thesis concerns the development of computational models and methods for the simultaneous inference of duplications and LGTs.

An evolutionary process on a much smaller scale is seen in cancer progression.

Cancer progresses via accumulation of genetic changes, and evolutionary mechanisms such as selection, competition, predation, and genetic drift characterize this process. The order in which different genetic changes appear during progression of the disease varies between distinct types of cancer. Although cancer progression is best described using graphs or networks, construction of such networks from available data is quite difficult. The last part of this thesis is concerned with a model of cancer progression based on trees, and mixtures thereof, for which we are able to develop efficient algorithms.

The outline of the rest of this thesis is as follows. Chapter 2 provides biological background on evolution and cancer progression. Also, a discussion on the use of trees and methods for tree reconstruction is provided. In Chapter 3, we introduce the major computational techniques that our methods are based on. Chapter 4 describes problems and methods for inference of duplications and lateral gene transfers. Section 4.5 describes the combinatorial model of gene evolution presented in Papers I and II, and Section 4.6 describes the probabilistic model of gene evolution presented in Paper III. Finally, Chapter 5 describes the methods previously developed for construction of cancer progression pathways, and the model and algorithms presented in Paper IV are discussed in Section 5.2. For a brief description of the articles included in this thesis, see Chapter 6.

A hole had just appeared in the Galaxy . . . Somewhere in the deeply remote past it seriously traumatized a small random group of atoms drifting through the empty sterility of space and made them cling together in the most extraordinarily unlikely patterns. These patterns quickly learnt to copy themselves (this was part of what was so extraordinary about the patterns) and went on to cause massive trouble on every planet they drifted on to. That was how life began in the Universe.

— Douglas Adams,
The Hitchhiker's Guide to the Galaxy

Chapter 2

Evolution and Describing its History

The underlying theme of the work presented in this thesis is molecular evolution and computational methods for its study. In this chapter we will give a brief background on and overview of molecular evolution in general, followed by a more detailed discussion of the evolutionary events at the focus of this thesis.

The outline of this chapter is as follows. Sections 2.1 and 2.2 give a brief overview of molecular evolution. In Section 2.5, we will discuss two important evolutionary events that have made a substantial impact in the genetic composition of organisms and which constitute a major focus of this thesis, namely, gene duplications and lateral gene transfers. The process of speciation, i.e., the evolutionary process in which new species emerge from existing ones, is discussed in Section 2.3. Trees have long been used as tools to depict the evolutionary history of organisms. Today they are also used to depict the history of genes that share a common ancestry, so-called homologous genes. A brief discussion of the use of trees and methods for their construction is given in Section 2.4.

A seemingly different, yet closely related, subject is that of cancer progression. “Cancer” is a name that refers to a large class of diseases with uncontrolled cell growth and proliferation as a common characteristic. The abnormal properties of cancerous cells are due to accumulation of harmful genetic changes. This process, called *somatic* evolution, is very similar to evolution of species and is discussed in Section 2.6.

2.1 Introduction to Genetics and Genomics

The observation that offspring inherit traits from their parents has long been used by humans, e.g., in breeding of animals and plants. Gregor Mendel performed the first systematic study of the basis of inheritance for some simple discrete traits, such as the color of the flower of the common pea plant [107, 108]. His discoveries concerning dominant and recessive traits became known as Mendelian inheritance.

Mendel's work did not receive any significant attention until it was rediscovered in the beginning of the 20th century. Research then led to the discovery that genes, the basic functional units of heredity, reside on the chromosomes, a discovery that was awarded with the Nobel prize in 1933. Although chromosomes were identified as the carriers of genetic material, the composition of the genetic material was yet unknown. The first experiments showing that the genetic information was contained in the DNA of chromosomes were performed by Avery *et al.* [8] and was later confirmed by Hershey and Chase [77].

James D. Watson and Francis Crick published the first accurate model of DNA structure in 1953 [150], and the genetic code was cracked by Har Gobind Khorana, Robert W. Holley, and Marshall Nirenberg, who shared the Nobel prize in physiology in 1968.

The central dogma of molecular biology, that the flow of information in the cell goes from DNA to mRNA, to protein, but never from protein to nucleic acid, was formulated by Francis Crick [56, 31].

Today, technological advances have enabled us to sequence entire genomes. All 6 billion bases comprising the genome of James D. Watson were sequenced in two months time in 2008 [152]. One of the challenges of the future lies in constructing computational tools for extracting functional information from sequence data. In the following, we give a brief overview of the molecular machinery of the cell responsible for reading the genome and producing the proteins that are responsible for most of a cell's functions. For more in-depth information, we refer to the standard text book by Bruce Alberts *et al.* [1].

The Central Dogma of Molecular Biology

The cell is the smallest structural and functional unit of an organism that is classified as living. There are two types of cells: eukaryotes comprising multicellular animals, plants, fungi, as well as unicellular organisms, and prokaryotes such as bacteria. "Karyose" comes from a Greek word meaning kernel, "pro" means before, and "eu" means true. So prokaryotic means "before a nucleus", and eukaryotic means "possessing a true nucleus". The name emphasizes the fact that eukaryotes carry their genetic material inside a cell nucleus, while prokaryotes have no such compartment and the genetic material is held within the cytosol. The genetic material of both eukaryotes and prokaryotes consists of long molecules of deoxyribonucleic acid (DNA). As the eukaryotic DNA molecules are very long and have to fit in a small nucleus, they are folded up into chromosomes in a highly organized manner. The prokaryotic DNA, on the other hand, is present as circular naked DNA molecules. DNA acts like an instruction manual and its sequence provides all the information needed for a cell to function. The information is first copied to ribonucleic acid (RNA) before being transformed into proteins—this is the so-called dogma of molecular biology, namely that information passes from nucleotides to amino acids but never in the opposite direction. The functional units in the DNA

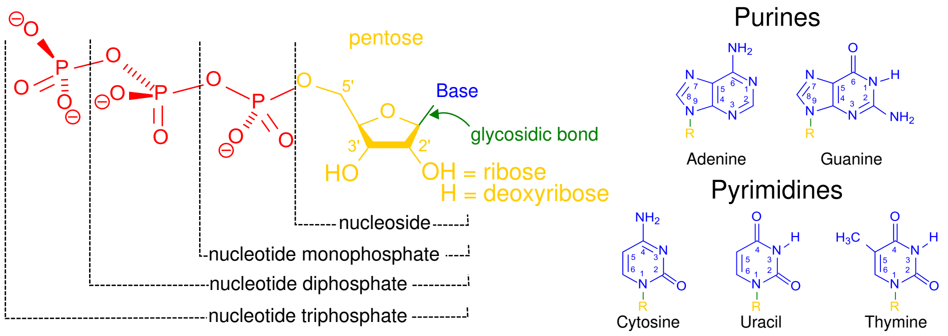

Figure 2.1: ***The structure of nucleotides.***

that code for RNA or proteins are called genes. Each gene encodes one or a set of similar proteins, and each protein performs a specialized function in the cell. Cells use the two-step process of transcription and translation to read genes and produce the strings of amino acids that make up a protein. The production of the various proteins is one of the most important processes occurring inside a cell as proteins not only form structural components of the cell, but they also compose the enzymes that catalyze the production of other organic biomolecules required for the cell to function.

DNA and RNA Structure

DNA is the carrier of genetic information composed of four different nucleotides. Each nucleotide is composed of three parts: (1) a nitrogenous base known as purine (adenine (A) and guanine (G)) or pyrimidine (cytosine (C) and thymine (T)); (2) a sugar, deoxyribose; and (3) a phosphate group. The nitrogenous base determines the identity of the nucleotide, and individual nucleotides are often referred to by their base (A, C, G, or T), see Figure 2.1. One DNA strand can consist of up to several hundred million nucleotides. The nucleotide T can form a hydrogen bond with A, and C with G, making a double-helix formed by two anti-parallel complementary strands, see Figure 2.2.

RNA is very similar to DNA, the only difference being that the pyrimidine base thymine is replaced by uracil (U) and the ribose comes in its fully hydroxylated form. Together, the presence of uracil in place of thymine, and the 2′-OH in the ribose constitute the two chemical differences between RNA and DNA. Also, RNA does not form a double helical structure and is in general single-stranded. There are many types of RNA present in the cell distinguished by their functional role. Three of these, namely messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), will be discussed in further detail below.

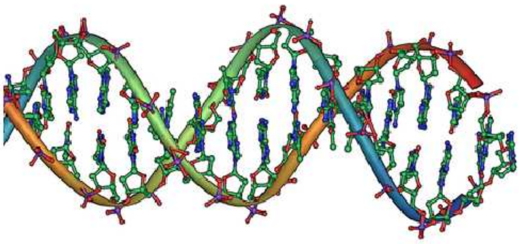

Figure 2.2: ***DNA double helix.***

Transcription

Transcription refers to the transfer of the genetic code from DNA to a complementary RNA and occurs in the cell nucleus. The mRNA serves as an intermediate between DNA and protein. The transcription starts with the enzyme RNA polymerase attaching and unzipping the DNA molecule into two separate strands after which it binds to the promoter segment of DNA that indicates the beginning of the single strand of DNA to be copied. It moves along the DNA and matches each DNA nucleotide with a complementary RNA nucleotide to create a new RNA molecule patterned after the DNA. The copying of the DNA continues until RNA polymerase reaches a termination signal, i.e., a specific set of nucleotides that mark the end of the gene to be copied. When the RNA polymerase has finished copying a particular segment of DNA, the DNA reconfigures into the original double-helix structure. In prokaryotes, this RNA needs no further processing and provides the blueprint which directs protein synthesis. However, in eukaryotes, this RNA strand (the transcript) is first processed into mature mRNA. The processing involves the removal of intervening non-coding sequences, so-called introns. The newly created mRNA is then exported out of the nucleus and into the cytoplasm where translation can take place.

Translation

Translation refers to the process of converting the information contained in an mRNA molecule into a sequence of amino acids that bind together to form a protein. In the cytosol, mRNA molecules bind to protein-RNA complexes called ribosomes. Each ribosome includes a large and a small subunit containing rRNA and more than 50 proteins. The small and large subunits of the ribosome surround the mRNA after which tRNA molecules carrying amino acids attach to the ribosome and mRNA to create the polypeptide chain, see Figure 2.3. There are several types of tRNA molecules each, containing a unique three base region called the anticodon that can base pair to the corresponding three base codon on the mRNA. Each type of tRNA

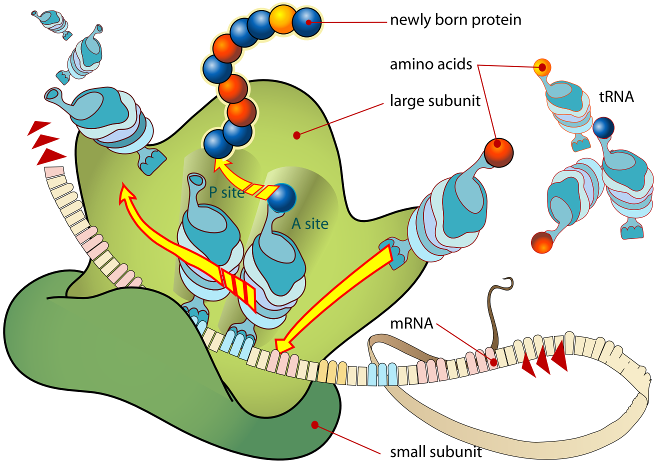

Figure 2.3: ***A schematic view of translation.***

molecule can only carry one unique amino acid, but may base pair to more than one codon sequence on the mRNA. Hence, each three base codon signals for the inclusion of a specific amino acid, but the same amino acid can be coded by several different codons. The correspondence between codons and amino acids is called the genetic code and is shown in Figure 2.4.

2.2 Genome Evolution

About 3.5 billion years ago, cells similar to modern day bacteria had appeared. There is evidence for the existence of eukaryotic cells 1.4 billion years ago with the first multicellular animals appearing around 640 million years ago [21]. In this section, we provide a few examples of the diverse (molecular) evolutionary events that have shaped present day genomes during millions of years of evolution.

Although cells have acquired highly complex and accurate mechanisms for DNA replication and repair, a cell can still fail to create exact copies of its chromosomal DNA during cell division. In fact, such failures are the predominant causes of genetic changes during evolution, although transposable DNA elements also play a major role.

In the context of this thesis we assume that there exists a reference genome that is representative of the genome of all individuals belonging to a certain species. For now, we also assume that the concept of species and the classification of individuals as belonging to one species is unproblematic, though as we will see, this has been contested in recent years mainly withing the prokaryotic domain.

We can imagine following the fate of a single gene as it is passed from one gen-

	2nd Position				
1st Position	**U**	**C**	**A**	**G**	3rd Position
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	*	*	**A**
	Leu	Ser	*	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

Figure 2.4: ***The genetic code.*** *A * indicates stop-codon.*

eration to the next in one species. For various reasons, we would observe changes to the DNA sequence of the gene. On a small scale, we would detect substitutions, insertions, and deletions involving a few nucleotides. Translocations, exon duplications, exon shuffling, and gene conversion are examples of other small scale events that alter the sequence of a gene.

Mathematical models of sequence evolution, so-called substitution models, have been proposed and are routinely used, e.g., to reconstruct trees showing the relationship between genes from different organisms. These models together with tree reconstruction methods are discussed in Section 2.4.

Taking genes to be atomic units, i.e., ignoring changes to the DNA sequence, we would see larger scale events that affect entire genes. For example, portions of chromosomes are sometimes duplicated or deleted and may affect a whole set of genes that reside on those segments. We would see how genes are lost, for example due to segmental deletions that remove the sequence or a part of the sequence from the genome altogether, or deleterious mutations that cause the silencing of the gene. The number of genes can also increase via events such as gene duplications (caused, e.g., by segmental duplications or reverse transcription), lateral gene transfers (the transfer of genetic material from one species or individual to another), or interstitial deletions (segmental deletions that do not include chromosomal endpoints; these may cause two genes to be fused together into one gene). On the largest scale, we have whole genome duplications that doubles the number of chromosomes and hence also the number of genes in a species, but is usually followed by massive gene losses.

The work presented in this thesis deals mainly with two of the evolutionary events mentioned above, namely gene duplications and lateral gene transfers. These are therefore discussed in more detail later in Section 2.5.

2.3 Speciations and Organismal Trees

Ever since Darwin's work popularized the idea of evolution, trees have been widely used to depict the historic relationship between species. When groups of individuals belonging to the same species are isolated from each other, they are independently affected by the evolutionary processes. Over time the groups will form distinct species, an event that we call speciation. Trees are often the best representation of the process of speciation in higher organism, although plants and fish are known to hybridize to form new species. In these cases, the speciation process is best represented by networks.

The classification of micro-organisms into species can sometimes be problematic. For example, it has become increasingly apparent that lateral gene transfers have played a major role in prokaryotic evolution, and some have argued that a species tree representing prokaryotic evolution, at least the evolution of certain groups of taxa, may not exist. Others argue that although lateral gene transfers have played a major role, the notion of species and species trees are still meaningful representations of the evolutionary history of prokaryotes.

2.4 Phylogeny and Tree Reconstruction

Before the emergence of the modern theory of molecular evolution, classification of species and inference of phylogenies were based on morphological data. With the modern understanding of the molecular mechanisms of inheritance, practice has shifted to using DNA or amino acid sequence data as the basis for reconstructing evolutionary histories. The history of a set of homologous genes is usually adequately represented by a tree, although there are some events such as gene conversion and recombination that are responsible for creating non tree-like histories. The history of species, whether represented by trees or networks, is often inferred from a set of gene trees. In this section we will discuss some of the more popular methods for gene tree reconstruction.

Computational methods for tree reconstruction attempt to find a tree or a set of trees that are optimal according to some criteria. Several different criteria have been used when devising computational methods for tree reconstruction. A number of methods have been proposed based on parsimony where we seek the tree that requires a minimum number of evolutionary events to explain the data (see Section 3.1 for a general discussion of parsimony). The first to suggest the use of parsimony as a criterion for tree reconstruction were Edwards and Cavalli-Sforza [45].

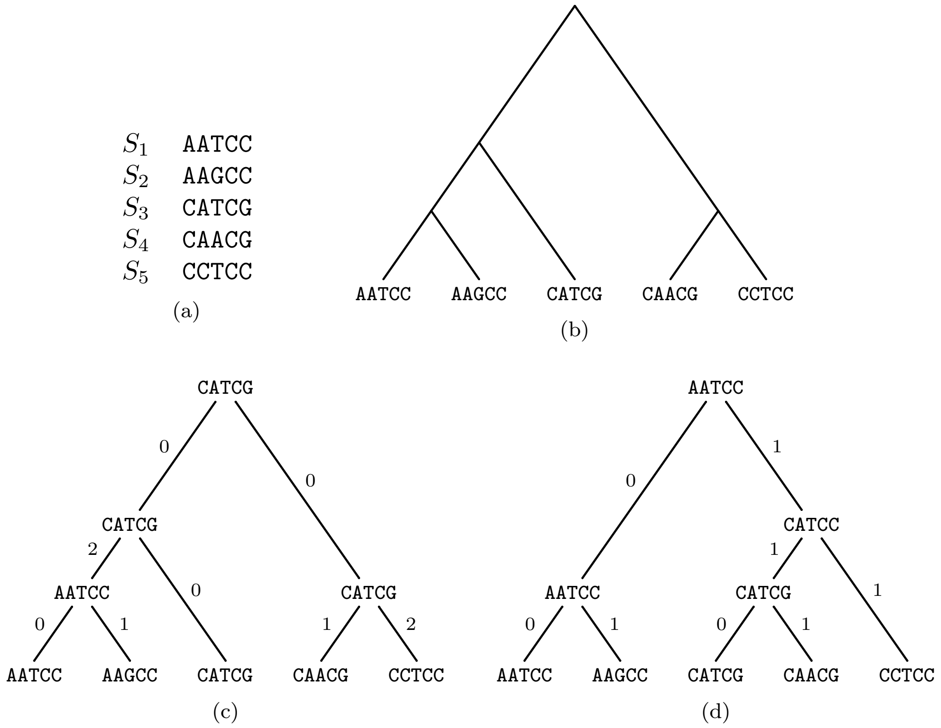

Figure 2.5: ***Maximum parsimony example**. (a) shows a set of five DNA sequences for which we seek a parsimony tree. (b) shows an arbitrary tree whose leaves are associated with the sequences. (c) shows one way to assign hypothetical sequences to the internal vertices to the tree in (b) such that the total number of substitutions is minimized. For this tree, at least 6 substitutions are required. (d) shows a maximum parsimony tree with hypothetical sequences assigned to internal vertices. This tree requires only 5 substitutions which is the least number of substitutions required for any tree and any assignment of sequences to internal vertices.*

Assume that we are given n sequences. For each rooted bifurcating tree with n leaves corresponding to the n sequences and an assignment of hypothetical sequences to the internal vertices, we can compute the minimal sets of evolutionary events that have taken place along each tree edge. Traditionally, an evolutionary event is defined as a single substitution of one nucleotide or amino acid for another. If we assign a cost to each possible substitution, we can compute a cost for each tree and assignment. The cost of a tree is then defined as the minimum cost over all possible assignments of sequences to its internal vertices. Out task is now to find the tree with the least cost. See Figure 2.5 for a complete example. A general algorithm for computing the minimal cost of a given tree was given by Sankoff [137] and Sankoff and Rousseau [138]. Finding the best tree can be done by searching in the space of trees using different heuristics and local search algorithms.

The criteria given above is quite general and various special cases together with specialized algorithms have been described in the literature. Other variants of parsimony exist that are not easily solved using the Sankoff algorithm, for example polymorphism parsimony [48, 51].

Distance matrix methods comprise another major class of phylogenetic methods. These methods were introduced in [25] and [58]. We can associate with each edge of a phylogenetic tree a branch length representing the total amount of evolution that has occurred between the two vertices. By summing the branch lengths of all the edges on the path between two leaves, we can compute pairwise distances between each pair of sequences or species. Distance matrix methods take as input pairwise distances between sequences and attempt to reconstruct a phylogenetic tree with branch lengths such that the distances induced by the tree is as close as possible to the given distances.

In order to develop algorithms for the problem, we need to define precisely how to measure closeness between the given pairwise distances and those induced by a tree. A popular, and statistically justifiable, criterion is *least squares.* Let D_{ij} be the given distance between sequence i and j and let D^T_{ij} be the distance as induced by a tree T. The best tree according to the least squares criterion is a tree T that minimizes the following expression:

$$\sum_{ij}(D_{ij} - D^T_{ij})^2. \tag{2.1}$$

There are efficient algorithms for determining branch lengths that minimize (2.1) given a fixed tree [63]. Finding the optimal tree according to the least squares criterion is, however, NP-hard [35]. Searching for a good tree is usually done by heuristics and local search methods.

Another criterion used by distance matrix methods is that of minimum evolution. In the minimum evolution methods, the branch lengths of a tree are determined by the least squares criterion, but the optimality criterion used to choose between trees is different. Instead of choosing the tree whose induced distances is as close as possible to the given distances, the tree with the minimum total length is preferred [87, 133].

There are also other heuristic distance matrix methods that do not have an explicit optimality criterion. One of the most popular methods for tree reconstruction is called neighbor-joining (NJ) [134] and is based on the clustering algorithms popularized by Sokal and Sneath [142]. Although NJ is defined by its algorithmic description, it is related to both the minimum evolution and least squares criteria. An interesting property of NJ is that it will reconstruct the correct tree if the given distances are "sufficiently close" to the distances induced by the tree. More formally, a distance matrix D is said to be *nearly additive* if there is a tree T with induced distance matrix D^T such that

$$|D - D^T_{ij}|_\infty < \frac{\mu(T)}{2},$$

where $\mu(T)$ is the minimum edge length in T. It can be shown that given a nearly additive distance matrix, NJ will reconstruct the unique tree T [6]. There are several other methods with this property that do not work as well in practice. However, FastNJ [47] is an algorithm that is very similar to NJ, works well in practice, and is much more efficient.

Distance matrix methods require calculations of pairwise distances of sequences. Due to insertions and deletions, homologous sequences are almost always of different lengths and the homologous positions are not immediately apparent. Hence, sequences must be *aligned* before computing distances. Many different exact and heuristic methods exist for sequence alignment and a proper discussion is beyond the scope of this thesis. We refer instead to several recent reviews on the state of modern alignment algorithms [91, 44, 123]. In general, alignment algorithms attempt to produce rows of sequences with inserted gaps such that the nucleotides or amino acids in each column are homologous. Often, columns that include gaps are discarded and distances are based only on columns without gaps.

Given an alignment, the naive approach to computing a distance between a pair of sequences is to count the number of mismatches. This leads, however, to an underestimation of the amount of substitutions that have occurred since more than one substitution may be responsible for a single mismatch or even a match since substitutions can be reversed by subsequent mutations. A better approach is instead to use a probabilistic model of sequence evolution to estimate the number of substitutions. Several substitution models have been proposed for DNA and amino acid sequences with varying amounts of complexity. Models of DNA evolution include Jukes-Cantor [84], Kimura's two-parameter model [88], F84 [54, 89], HKY [72], and the Tamura-Nei model [144].

Jukes-Cantor is the simplest model and assumes that substitutions occur according to a Poisson process with rate α and that all substitutions are equally likely. For this simple model, there are closed formulas for calculating the maximum likelihood estimate of the number of mutations. More complex models allow different rates to be assigned to different types of substitutions, for example transitions (substitution of one purine for another or of one pyrimidine for another) and transversions (substitution of a purine for a pyrimidine or vice versa). Numerical methods are used for the more complex models to obtain maximum likelihood estimates of distances or branch lengths. Substitution models exist also for protein sequences though we will not discuss them here.

Instead of using the substitution models described above to compute distances and then use distance methods to obtain a tree, it is possible to use the models and sequences more directly. Sequences together with a substitution model induce a probability distribution on trees with branch lengths. Hence, a natural problem is to find the maximum likelihood tree, i.e., the tree which maximizes the probability of observing the sequences given the substitution model. Phylogenetic likelihood methods were popularized by Felsenstein, see for example [52]. See also Section 3.3 which discusses structural EM algorithms.

Recent years have seen a growing body of Bayesian methods being developed.

Bayesian statistical inference is discussed in Section 3.4. A very popular computer program for Bayesian inference of phylogenetic trees is MrBayes [78, 132]. For an excellent introduction to phylogeny inference in general, see [53]. Felsenstein maintains a comprehensive list of phylogeny software that can be found at http://evolution.gs.washington.edu/phylip/software.html.

The discussion above has centered around construction of trees from sequence data. We often find that trees constructed from sequences of different genes from the same set of organisms are not identical. Species trees can be constructed based on sets of core genes, such as those involved in the transcription or translation machinery, whose evolution is believed to closely follow that of the corresponding species. Other methods include tree consensus methods that are based on the collective signal from a large set of gene trees, and the use of concatenated sequences from many different genes. For some groups of organisms, such as humans, apes, and rodents, we may have other kinds of information, such as archaeological data, available that can also be used for estimation of organismal phylogenies.

Irrespective of our method of choice, we are bound to observe that gene trees and species tree are different and that the evolutionary history of genes do not always follow that of the corresponding species. This poses the problem of reconciling the differences between trees by identifying the responsible evolutionary events. Tree reconciliation is a major part of this thesis and will be discussed in coming sections and chapters.

2.5 Gene Duplications and Lateral Gene Transfers

In this chapter, we will take a deeper look at the two evolutionary events with which this thesis is mainly concerned. The importance of gene duplication and its role as a major driving force of evolution has been established. This is in contrast to the role of lateral gene transfer (LGT) which has been the subject of much controversy. The next two subsections will deal with gene duplications and the controversy surrounding LGTs.

Gene Duplications

The role of gene duplication as a major driving force of evolution has been recognized for a long time. Ohno's seminal book *Evolution by Gene Duplication* [127] in 1970 popularized the idea among biologists, although it had been discussed and debated much earlier. For example, already in 1918, Bridges speculated that duplicate genes can mutate separately thus diversifying their functions [18]. See also later papers by Bridges [19] and Muller [117, 118]. For a review of the history of these ideas, see [145].

Several mechanisms are responsible for creating copies of a gene. These include unequal crossing-over, retrotransposition, segmental duplication, and whole genome duplication.

Unequal crossing-over occurs when homologous chromosomes are not precisely paired during recombination and results in chromosomes of unequal length: one chromosome acquires more genetic material than it passes over and thus contains a duplicated segment. This segment may contain part of a gene, an entire gene, or several genes. Genes duplicated via unequal crossing-over are located on the same chromosome, at least initially, before other events change their relative locations.

Retrotransposition is the process during which a messenger RNA (mRNA) is retrotranscribed to copy DNA (cDNA) and is then inserted into the genome, probably at a random location on some chromosome. The two versions of the gene generally reside on different chromosomes, and the copy also lacks introns since the introns have been spliced out before the mRNA is copied to cDNA.

Segmental duplications, i.e., duplications of large segments of a chromosome, are also responsible for duplication of genes and have been shown to have occurred frequently during primate evolution. The sizes of the duplicated segments tend to be somewhere between 1 000 to 200 000 nucleotides [135, 106]. The exact mechanisms creating such duplications are not clear though several models have been proposed [10].

Whole genome duplications, which may occur during abnormal cell division, is most commonly found in plants, and also in fishes [112]. There is also evidence indicating that one or two whole genome duplications have occurred very early in vertebrate evolution [36]. Whole genome duplications are usually accompanied by massive gene losses [29, 22].

The rate with which gene duplication occurs in eukaryotes has been estimated to one duplication per gene per 100 million years, which is similar to the rate of nucleotide substitutions [100]. Although the rate of duplication is high, most duplications are followed by gene loss. The fates of recently duplicated genes were termed non-functionalization, sub-functionalization, and neo-functionalization in [101, 59].

Analysis of sequenced genomes have revealed that a substantial proportion of genes are duplicated and that the distribution of gene family sizes across species varies greatly [154]. For example, the biggest gene family in *Drosophila melanogaster* has 111 members, while the biggest family in mammals is the olfactory receptor family with more that a thousand members. The KRAB-zinc finger family is another example of a gene family that has undergone many recent gene duplication events and there are over 400 active members of the gene present in the human genome [70].

Lateral Gene Transfers

Contrary to gene duplications, the importance and prevalence of lateral gene transfers has been much more controversial. LGT refers to the transfer of genetic material from one individual to another. The possibility of LGT in bacteria was realized already in the 1940s [93, 94] and demonstrated to occur between different species in 1959 [125]. We know today that LGT occurs frequently among prokaryotes [126, 20], and that it also occurs from prokaryotes to eukaryotes and among eukaryotes [85], though probably not as frequently as in the prokaryotic domain.

The abundance of LGTs in prokaryotes has led to some researchers challenging the idea that phylogenetic trees are able to represent prokaryotic evolution, see for example [64, 41, 153], and also [42] and references therein. There is an emerging view today that although LGTs have occurred among prokaryotes, perhaps it has not occurred so much that we must abandon trees altogether [11].

The mechanisms of LGTs among prokaryotes include transformation, transduction, and conjugation. Transformation refers to the uptake of DNA which is then incorporated into the genome. Certain bacteria have a natural ability to take up DNA from their environments. Transduction refers to the process of genetic exchange between bacteria mediated by a bacterial virus, a bacteriophage. Conjugation is a process in which bacterial cells transfer genetic material via direct contact.

Lateral gene transfer has also become an important medical issue [143] as it plays a major role in the spread of antibiotic resistance genes among pathogenic bacteria. Recently, the role of LGT in pathogen evolution has received much attention [92].

2.6 Progression in Cancer—an Evolutionary Phenomenon

Cancer is the name given to a whole host of genetic diseases in which cells undergo uncontrolled growth. Genetic alterations to three types of genes are responsible for tumorigenesis—the process in which normal cells are transformed into cancer cells. These are the gatekeepers, caretakers, and landscapers [113]. Gatekeepers are genes that directly affect growth and differentiation of cells and include the oncogenes and tumor suppressor genes, i.e., genes whose abnormal activation and suppression, respectively, can turn normal cells into cancer cells. Caretakers are responsible for maintaining the genomic integrity of cells and promote tumorigenesis indirectly. Alterations to caretaker genes can lead to genetic instability causing rapid accumulation of changes to the genome. Such changes can affect oncogenes or tumor suppressor genes which in turn leads to abnormal proliferation. As the name suggests, landscapers affect the micro-environment of cells. Landscaper genes cause tumorigenesis indirectly by generating an abnormal stromal environment [15]. When the normal intercellular signals are disrupted, for example during sustained inflammation, cells possessing tumorigenic potential can start to proliferate uncontrollably. Such abnormal conditions can also cause genetic instability which then could lead to development of cancer. For example, it has been known for more than a century that inflammation associated with tissue wounding can produce tumors, see [40, 140] and references therein.

Although cancer is a generic name for different diseases, six “hallmarks of cancer” common to all cancer types have been identified [71]. These are self-sufficiency in growth signals, insensitivity to anti-growth signals, apoptosis-evasion (evasion of programmed cell death), limitless replicative potential, sustained angiogenesis (the growth of new blood vessels), and tissue invasion and metastasis. Acquiring all these traits requires major genetic alterations that are accumulated as cancer progresses towards further malignancy. Although the rate of nucleotide mutation does not appear to be higher in cancer cells [148], chromosomal instability (CIN) seems to be present in all types of human cancer [96]. Mutations in CIN genes increase the rate with which whole chromosomes or large parts of chromosomes are lost or gained during cell division. Aneuploidy, i.e., imbalances in the number of chromosomes, and increased rates of loss of heterozygocity are caused by CIN. In [71], CIN was not identified as a hallmark of cancer but was taken to be a prerequisite for acquiring the entire set of hallmarks.

The progression of cancer and the acquisition of the previously mentioned traits is an evolutionary process involving selection among genetically variable cells [124, 30]. The evolution of cells within the body is called somatic evolution. Somatic evolution shares many similarities with evolution of organisms and many methods and models from population genetics [113] and ecology [30] can be applied to cancer progression, although some differences do exist and models may have to be altered to take these into consideration. Important factors that play major roles in somatic evolution include mutation, genetic drift, natural selection, competition, predation, and dispersal or colonization. A neoplasm, or tumor, consists of a large population of genetically heterogeneous individuals [103, 24] that undergo selective sweeps followed by clonal expansions, see for example [105] and [104]. Clones, i.e., a group of cells derived from a single mother cell, can expand or contract based on their fitness in the population. In general, evolution within a tumor population selects for increased growth and survival and mutations can become fixed in the population during selective sweeps.

The rates of mutation in cancer cells remain undetermined *in vivo*, but there are indications that they are not higher than in normal cells. In cell culture, the rates have been determined to be somewhere between 10^6 and 10^7 per locus per generation [2]. In [148], the number of non-synonymous mutations in colorectal tumor cells was determined to approximately 1 mutation per megabase of DNA, which is similar to the expected number of mutations in normal cells that have undergone as many generations and population size bottlenecks. The number of mutations required to cause cancer is not precisely known, but is probably somewhere between 3 and 12 depending on the type of cancer [131]. Given the incidence rate of cancer in the human population, it seems unlikely that so many mutations could be accumulated in a single cell based solely on normal somatic mutation rates [99, 71]. One explanation could be that the expansion of clones provides large enough populations to produce subsequent necessary mutations [115].

In some cases mutation in a gene gives its host a selective advantage only after another gene has undergone mutation. For example, in Barrett's Esophagus, the inactivation of *TP53* almost always occurs after *CDKN2A* has been inactivated [105]. It could be the case that inactivation of *CDKN2A* initiates a clonal expansion that provides opportunities for mutations to occur to *TP53* after which a second clonal expansion would occur. Such dependencies provide opportunities for modeling of cancer progression which is discussed in Chapter 5.

Competition and predation are other evolutionary forces that can act on organisms and have analogs in somatic evolution. For example, cells in a tumor are constantly competing for resources. More complex modes of competition have also been shown. For example, clones on different locations in the same mouse or rat can inhibit each other's growth [114, 23]. Aspects similar to predation in ecology is also present in cancer progression, most naturally from the immune system. One difference compared to ecology is that the extinction of prey does not lead to the extinction of the predator.

Chapter 3

Computational Techniques

This chapter provides some background on a selection of computational techniques that have been used in the work presented in this thesis. Parsimony can be explained as a general principle of "less is more". Papers I and II deal with methods for finding the "simplest" reconciliations of trees and Section 3.1 provides a brief overview of the subject of parsimony. In Paper I, we also develop a fixed-parameter tractable algorithm for the tree reconciliation problem. Some general comments about parametrized complexity is given in Section 3.2. Expectation Maximization (EM) is an iterative meta algorithm for finding maximum likelihood estimates in probabilistic models. Paper IV of this thesis provides an EM algorithm for a model of cancer progression. A general description of EM algorithms is given in Section 3.3. Bayesian methods have been applied quite successfully on a large set of problems in bioinformatics, though their use is sometimes controversial. In Paper III of this thesis, we develop algorithms and methods for Bayesian inference of duplications and lateral gene transfers in which Markov Chain Monte Carlo (MCMC) techniques play a major role. A brief overview of Bayesian methods and the degree-of-belief interpretation of probability is given in Section 3.4 along with a discussion of Markov Chain Monte Carlo (MCMC) Techniques.

3.1 Ockham's Razor and Parsimony

Numquam ponenda est pluralitas
sine necessitate
William of Ockham

In bioinformatics, maximum parsimony is probably the most well-known example of the use of the principle of Ockham's razor—that "plurality should never be posited without necessity". In [45], Edwards and Cavalli-Sforza first mentioned the general idea of maximum parsimony when they declared that the preferred evolutionary tree is the one that involves the minimum net amount of evolution, and for a long time, parsimony methods were the most widely used tree reconstruction methods for character data.

The general principle of parsimony is, of course, not restricted to evolutionary trees. In fact, the principle has been advocated many times in the past, even long before Ockham applied it to such an extent as to give it its current name. A famous example in scientific literature is Newton's first rule of reasoning in philosophy as stated in "Mathematical Principles of Natural Philosophy":

> *Rule I We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.*
>
> *To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.*
>
> *RULE II Therefore to the same natural effects we must, as far as possible, assign the same causes.*

In general, parsimony can be applied in two different ways. On the one hand, we can take parsimony as an optimality criterion, as in the case of maximum parsimony in phylogenetics: trees are scored based on the level of complexity with which they can explain the data and the trees which require fewer assumptions of evolutionary events are preferred over trees that require more. On the other hand, parsimony can be applied when designing models in the sense of defining models with no more parameters than necessary to sufficiently model the data. A simple example of this is fitting a curve to points in the plane. In finance, models for prediction of yield curves are another example where parsimony has been explicitly applied when devising models [121, 39]. Many more examples can be found where parsimony is used implicitly, and there is a vast amount of literature in statistics concerning the choice of model and the number of parameters.

In this thesis, parsimony is applied to the problem of tree reconciliation which is explained in Chapter 4. Analogous to tree reconstruction, parsimony is here used as an optimality criterion where the simplest reconciliation is sought in the sense of minimizing the number of evolutionary events needed to explain the differences between an organismal tree and a corresponding gene tree. Early work on tree reconciliation problems sought to reconcile trees using duplications and losses. In this case, defining the underlying combinatorial model is quite straightforward. A level of complexity is added when we also have to consider lateral gene transfers, and care must be taken when defining the combinatorial problem to ensure biological feasibility. The combinatorial model developed in this thesis is biologically feasible and an extensive discussion of this issue can be found in Paper II.

A common objection to the use of parsimony is that nature and evolution are not constrained to being parsimonious. The usual answer is that we are not making an assumption about nature or evolution, and that parsimony methods can yield quite complex solutions. We merely strive to find a minimal set of assumptions needed to explain the data. In any case, the strength of any method lies in its predictive strength, or in the case of tree reconstruction, the ability to correctly infer the past. Even when the true process that generates the data is highly complex, it may be a good idea to use simple models, at least when the sample size is small.

3.2 Parameterized Complexity

Computational complexity theory is a branch of theoretical computer science concerned with analyzing the amount of resources needed to solve computational problems. The most important computational resources are time and memory. In complexity theory, problems are categorized into complexity classes based on the amount of resources needed to solve them. The classes P and NP are the most studied due to their practical implications, and the question $\mathrm{P} \overset{?}{=} \mathrm{NP}$ is one of the most important unsolved problems in theoretical computer science and mathematics. For an introduction to this topic, we refer the reader to the classic book of Garey and Johnson [62].

The time complexity of an algorithm is measured as a function, say f, of the input size. If n is the size of the input, then $f(n)$ is the maximum number of steps that the algorithm needs to produce its output. Under the unit cost model, a step is any basic operation such as addition, multiplication, or comparison. Although there are cases when it is preferable to keep separate counts of different operations, e.g., by analyzing the number of multiplications and additions separately, we will only consider the unit cost model in this text. When comparing the time complexity of different algorithms, we are mainly interested in their asymptotic behavior. We say that a function $f(n)$ is $O(g(n))$, if there is a constant C such that $|f(n)| \leq C \cdot |g(n)|$ for all sufficiently large n. A polynomial time algorithm is one whose time complexity is $O(p(n))$ where p is some polynomial. Algorithms with polynomial time complexity are considered efficient and problems for which polynomial time algorithms are known are considered tractable. This definition of computational efficiency has proved extremely successful when dealing with natural or real-world problems; in the vast majority of real-world cases, polynomial time algorithms are sufficiently efficient.

When confronted with a problem that does not seem to admit a polynomial time solution, the traditional way to deal with it is to try to show that the problem is NP-hard. The theory of P and NP deals only with decision problems. As an example, take the traveling salesman problem in which we are given a set of cities together with distances between each pair. We want to find a minimal length tour that visits all cities. This optimization problem can be recast as a decision problem: given the cities and the distances between them, is there a tour that visits all cities and whose total length is at most K? Clearly, the optimization problem is at least as hard as the decision problem, and so, if the decision problem can be shown to be hard, then the optimization problem must also be hard. The complexity class P consists of all decision problems that admit a polynomial time algorithm. The class NP consists of all decision problems whose "yes"-instances can be verified in polynomial time. By *verifying the "yes"-instances*, we mean that for each "yes"-instance, there is a certificate with the help of which we can check that the instance really is a "yes"-instance. For example, in the case of the traveling salesman, a certificate consists of a minimal length tour, and checking that a tour visits all cities and has length no more than K can be done in polynomial time. Hence, traveling salesman is in NP. In fact, it is one of the most well-known examples of NP-complete problems. From the definition of P and NP, it is clear that P is a subset of NP. Most researchers believe that the converse is not true, although no proof of this fact has been found.

One benefit of studying decision problems is that we can describe them in terms of the formal notion of languages. For any finite non-empty set of symbols Σ, let Σ^* denote the set of all strings of symbols from Σ. A set L is a language over the alphabet Σ if it is a subset of Σ^*. The instances of a decision problem can always be encoded by strings of symbols from Σ, e.g., when $\Sigma = \{0, 1\}$. The language corresponding to a decision problem $\mathcal{P}$ is simply the subset $L_{\mathcal{P}} \subseteq \Sigma^*$ whose members are the encoded "yes"-instances of $\mathcal{P}$. We say that an algorithm decides a language L, if it returns "yes" when presented with an element of L and "no" otherwise. A problem $\mathcal{P}$ is in the class P if there is a polynomial time algorithm that decides $L_{\mathcal{P}}$.

A problem is NP-complete if every problem in NP can be reduced to it via a polynomial time algorithm. We say that a problem $\mathcal{P}_1$ can be reduced to problem $\mathcal{P}_2$ if there is a polynomial time algorithm that transforms each instance x_1 of $\mathcal{P}_1$ into an instance x_2 of $\mathcal{P}_2$ such that x_1 is a "yes"-instance of $\mathcal{P}_1$ if and only if x_2 is a "yes"-instance of $\mathcal{P}_2$. Clearly, if $\mathcal{P}_2$ can be solved in polynomial time, then so can $\mathcal{P}_1$.

All hope is not lost when a problem is shown to be NP-complete. The set of NP-complete problems consist of many crucial real-world problems that need to be solved despite being hard. For example, many heuristics exist for various optimization problems that work well in practice, at least for certain subsets of the problem instances. Sometimes optimization problems admit approximation algorithms with guaranteed performance. See for example [7] for general discussions on approximation algorithms and complexity classes. In the majority of cases, however, the naive brute force algorithms do not work well in practice. Consider for example another famous NP-complete problem, namely vertex cover:

Vertex Cover
Instance: A graph G and a non-negative integer $K \leq |V(G)|$
Question: Is there a set of vertices $V \subseteq V(G)$ of size K such that V covers G?

A set V of vertices is said to cover G if each edge of G is incident to at least one vertex in V. The brute-force algorithm for this problem simply consists of checking every vertex subset of size K. There are $O(n^K)$ such subsets and as n becomes large, checking them all, even for small k is infeasible.

In 1986, Fellows and Langston [50], observed that vertex cover could be solved in time $O(f(K)n^3)$. Later a very simple and elegant algorithm was discovered that runs in time $O(2^K n)$ [49]. Note how this time complexity separates the size of the input from the parameter K. The implication is that the algorithm is polynomial in the size of the input, and exponential only in the parameter. Hence, the problem is tractable even for large instances, as long as the minimal cover set is small. Improvements have since been made for vertex cover, and algorithms being able to handle K up to about 400 have been implemented and used in multiple sequence alignment problems [27]. This is an example of parameterization of time complexity and the algorithm mentioned above for vertex cover is called a fixed-parameter tractable algorithm.

More formally, a paramterized problem is a subset of $\Sigma^* \times \mathbb{N}$. An instance of a paramterized problem is a pair (I, K), where K is the so-called parameter. The run-time of an algorithm for a parameterized problem is a function of $|I|$ and K. A parameterized problem is said to be fixed-parameter tractable (FPT) if there exists an algorithm for the problem with time complexity $O(f(K) \cdot |I|^c)$, where c is a fixed constant and f is a function of K that does not depend on I. The parameterized version of vertex cover can be stated as follows:

k-Vertex Cover
Instance: A graph G and a non-negative integer $K \leq |V(G)|$
Parameter: K
Question: Is there a set of vertices $V \subseteq V(G)$ of size K such that V covers G?

We note, in conclusion, that a problem may have many possible parameterizations. A problem can be fixed-parameter tractable for some parameterizations and not so for others. There is also a hierarchy of complexity classes in the theory of parameterized complexity. For a thorough treatment of this subject, we refer the interested reader to [43].

3.3 Maximum Likelihood Estimation with Expectation Maximization

Classic statistical inference can be divided into parametric and non-parametric. In non-parametric inference, no specific type of probability distribution or model is assumed. Instead, other kinds of hypotheses are made, for example, a common assumption is that the data are observations of independent and identically distributed (iid) random variables. It is, in general, quite difficult to incorporate previous knowledge or beliefs about the underlying real-world structure that has generated the data in non-parametric inference methods. In parametric inference, the observed data are assumed to be observations from some family of probability distributions. Examples of such distributions include the normal or Gaussian distribution, the multinomial distribution, and the Dirichlet distribution. Distributions may also be specified using probabilistic (generative) models containing structural elements attempting to capture the most important features of the real-world situation that has generated the data. An example of such a distribution family is phylogenetic trees that we can think of as generating sequences. In any case, a distribution belonging to a family is determined by a set of parameters. For the normal distribution the parameters are the mean and variance, whereas for phylogenetic trees, parameters include the tree topology and the parameters of our chosen sequence evolution model (see Section 2.4). The classical inference problem is then to find, or estimate, the set of parameters that best fit the data according to some criteria.

The most widely used criteria for estimating the parameters is probably maximum likelihood (ML). The likelihood of a parameter set θ is simply the probability of the observed data X given the parameters:

$$L(\theta|X) = p[X|\theta].$$

The maximum likelihood estimate of θ is defined as the θ^* maximizing the likelihood, or more generally, any function that is proportional to the likelihood. Note that the likelihood function is really a function of θ alone since we regard the data as fixed, and is not a probability distribution, i.e., in general $\int_\theta L(\theta|x) \neq 1$.

It is sometimes possible to determine the ML estimate by deriving closed formulas, but in many cases such a method is infeasible. A popular computational technique for parameter estimation is Expectation Maximization (EM). The method has been in use in different forms for a long time, but was generalized and popularized with the publication of a paper by Dempster, Laird, and Rubin in 1977 [37]. For some notes on the history of the EM algorithm, see [109].

EM has been successfully applied to a wide variety of problems in diverse scientific fields and many modifications and improvements have been suggested, see for example [109, 79, 110, 55]. In the next subsection, we will give the standard derivation of the EM algorithm together with a proof of its convergence. Subsequently, we will discuss the structural EM algorithm of Friedman *et al.* [60, 61] which is directly related to the work presented in Paper IV of this thesis.

Standard EM

Let $X = \{x_1, \ldots, x_N\}$ be the set of observed data, and let θ denote the set of parameters. In many applications, we also have a set of missing data or hidden variables. These are sometimes introduced in the model in order to simplify computations of certain probabilities. As a concrete example, assume that X consists of a set of points on the real line and that we wish to model the data using a mixture Y of two normal distributions:

$$\begin{aligned}
Y_1 &\sim N(\mu_1, \sigma_1^2), \\
Y_2 &\sim N(\mu_2, \sigma_2^2), \\
Y &= \begin{cases} Y_1 & \text{with probability } \pi, \\ Y_2 & \text{with probability } 1 - \pi. \end{cases}
\end{aligned}$$

Thinking of the model as generative, the above notation has the following interpretation: each data point is generated from distribution Y_1 with probability π or Y_2 with probability $1 - \pi$. In this case, θ consists of five parameters:

$$\theta = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2).$$

The computation of certain probabilities are made easier by introducing a set Z of hidden variables indicating the distribution from which each data point in X was generated:

$$z_i = \begin{cases} 0 & \text{if } x_i \text{ was generated from distribution } Y_i, \\ 1 & \text{otherwise.} \end{cases}$$

The idea of the EM algorithm is as follows. At the start of the nth iteration we have a set of parameters θ_n. Together with the observed data X, θ_n induces a probability distribution on the hidden variables, $\Pr[Z|X, \theta_n]$. The set of parameters θ_{n+1} for the next iteration is obtained by finding the θ that maximizes the expectation of the so-called complete-data log-likelihood

$$E_{Z|X,\theta_n}\big[\log p[X, Z|\theta]\big],$$

where the expectation is taken over the distribution $\Pr[Z|X, \theta_n]$. The procedure is guaranteed not to decrease the likelihood, in other words $L(\theta_{n+1}|X) \geq L(\theta_n|X)$. We next show why this actually works.

First, we show that the likelihood can be written as the sum of two expectations:

$$\begin{aligned} \log L(\theta|X) &= \log p[X|\theta] \\ &= \log p[X|\theta] \cdot 1 \\ &= \log p[X|\theta] \left(\sum_Z \Pr[Z|X, \theta] \right) \\ &= \sum_Z \Pr[Z|X, \theta] \log p[X|\theta] \\ &= \sum_Z \Pr[Z|X, \theta] \log \frac{p[X, Z|\theta]}{\Pr[Z|X, \theta]} \\ &= \sum_Z \Pr[Z|X, \theta] \log p[X, Z|\theta] - \sum_Z \Pr[Z|X, \theta] \log \Pr[Z|X, \theta] \\ &= E_{Z|X,\theta}\big[\log p[X, Z|\theta]\big] - E_{Z|X,\theta}\big[\log \Pr[X|Z, \theta]\big]. \end{aligned} \tag{3.1}$$

For the next result, we use Jensen's inequality to obtain

$$\begin{aligned}
\log L(\theta|X) &= \log p[X|\theta] \\
&= \log \sum_{Z} p[X, Z|\theta] \\
&= \log \sum_{Z} \Pr[Z|X, \theta'] \frac{p[X, Z|\theta]}{\Pr[Z|X, \theta']} \\
&= \log E_{Z|X,\theta'} \left[\frac{p[X, Z|\theta]}{\Pr[Z|X, \theta']} \right] \\
&\geq E_{Z|X,\theta'} \log \left[\frac{p[X, Z|\theta]}{\Pr[Z|X, \theta']} \right] \qquad \text{(By Jensen's inequality)} \\
&= E_{Z|X,\theta'} \big[\log p[X, Z|\theta] \big] - E_{Z|X,\theta'} \big[\log \Pr[X|Z, \theta'] \big]. \qquad (3.2)
\end{aligned}$$

Noting the similarity between (3.1) and (3.2), we define the famous Q- and R-terms as

$$\begin{aligned}
Q(\theta, \theta') &= E_{Z|X,\theta'} \big[\log p[X, Z|\theta] \big] \\
R(\theta) &= -E_{Z|X,\theta} \big[\log \Pr[X|Z, \theta] \big]
\end{aligned}$$

Assume that we have a set of parameters θ_n at the start of the nth iteration and let

$$\theta^* = \operatorname*{argmax}_{\theta} Q(\theta, \theta_n). \qquad (3.3)$$

We now have that

$$\begin{aligned}
\log L(X|\theta^*) &\geq Q(\theta^*, \theta_n) + R(\theta_n) && \text{(By (3.2))} \\
&\geq Q(\theta_n, \theta_n) + R(\theta_n) && \text{(By definition of } \theta^*\text{)} \\
&= L(X|\theta_n). && \text{(By (3.1))}
\end{aligned}$$

Hence, by choosing $\theta_{n+1} = \theta^*$, we are guaranteed not to decrease the likelihood. All we need to do to implement an EM algorithm is to perform the maximization of the Q-term in (3.3). The maximization of the Q-term is generally done in two steps: In the E-step, certain quantities are computed that only depend on the current set of parameters and the observed data. This is in preparation for the M-step where the computed quantities are used to find the set of parameters that maximizes the Q-term.

We note here that maximization of the Q-term is not necessary for convergence. Convergence to a local optimum is guaranteed as long as we are able to find a set of parameters θ_{n+1} such that $Q(\theta_{n+1}, \theta_n) \geq Q(\theta_n, \theta_n)$. This procedure is called Generalized EM (GEM) and can be used when maximization of the Q-term is infeasible. One drawback of GEM compared to standard EM is a potentially slower rate of convergence.

In conclusion, both standard EM and generalized EM may suffer from the same drawbacks as local search methods, and finding a globally optimal solution may require different heuristics such as using a set of random start values or simulated annealing strategies.

Structural EM

In 1997, Friedman devised a structural EM algorithm for Bayesian networks, that beside improving the numeric parameters in each step, also improves the structure [60]. In 2002, the same approach was used for tree reconstruction [61].

Likelihood-based methods for tree reconstruction have been very successful and are quite popular. Prior to Friedman's contribution, methods for ML estimation of phylogenies used the EM algorithm only for optimization of the parameters on a fixed tree. When searching for the best topology, each tree considered would have to be passed to the EM algorithm for parameter optimization. This is computationally very expensive, and in practice, only a few selected topologies could be considered. Friedman *et al.* observed that it is possible to simultaneously improve the topology and parameters.

In this setting, the input consists of aligned sequences X, and the set of parameters of the model consist of both the tree topology T and the lengths l of the tree edges, i.e., $\theta = (T, l)$. Just as in standard EM, the parameters θ_n of the previous iteration induce a distribution on the space of topologies and lengths. The crucial observation made by Friedman *et al.* is that the contribution to the Q-term from each *possible* edge is the same for all trees and can be computed once and for all. These contributions are then used as weights on the set of all pairs of vertices and the problem of finding the best topology given θ_n is reduced to finding the bifurcating tree with greatest total weight. Unfortunately, this problem turns out to be NP-complete. To overcome this difficulty, Friedman *et al.* use the maximum spanning tree algorithm to obtain a tree that is not necessarily bifurcating, but which is then transformed into a bifurcating tree via modifying steps that are guaranteed not to decrease the likelihood. Hence, in each iteration of the EM algorithm, both the topology and the parameters are changed, leading to great savings in computational time.

We note here that an important distinction can be made among structural EM algorithms. Just as for standard EM and generalized EM, we can distinguish between structural EM algorithms that respectively, maximize and merely improve on the Q-term. When possible, an EM algorithm that maximizes the Q-term in each iteration is preferred due to faster convergence rates. In Paper IV of this thesis, we provide a structural EM algorithm that maximizes the Q-term in each step. In order to emphasize the distinction between maximizing and improving the Q-term, we call our algorithm a *global* structural EM algorithm.

3.4 Bayesian Inference with Markov Chain Monte Carlo

Bayesian statistical inference has become increasingly popular in the field of bioinformatics. In this section, we will give a brief background on Bayesian statistics and touch on some of the controversial issues. Finally, we will provide a discussion on MCMC techniques that is relevant to the work presented in Paper III of this thesis.

A Philosophical Question

Discussions of Bayesian versus classical statistics usually start with a philosophical question: what is a probability? The so-called frequentist answer is that the probability of an event A is the long run proportion of times that event A occurs during a large number of replications of an experiment. Hence, the probability of heads in a coin tossing experiment with a fair coin is 0.5 and the probability of obtaining a six on a roll of a single fair die is $\frac{1}{6}$. In contrast, the Bayesian answer is that probability is a measure of an individual's uncertainty about the outcome of an experiment, with the added constraint that the individual's opinion must be consistent, in other words, assignment of probabilities to events must be in accordance with the Kolmogorov axioms of probability theory. In the case of a toss of a coin and a roll of a die above, most Bayesians too would assign probabilities 0.5 and $\frac{1}{6}$ to the events of heads and six, respectively. However, to a Bayesian, *any* event can be assigned a probability as a measure of uncertainty, even if the experiment could never be replicated. More importantly, a Bayesian may assign a probability to a hypothesis, while to a frequentist, a hypothesis should either be rejected or retained. A highly cited and enjoyable classic paper on the subject of Bayesian subjective probability was written by Edwards, Lindman, and Savage [46].

Taking an example from [46], imagine a great prize being offered to predict the outcome of a coin toss. With no other previous knowledge, both the frequentist and the Bayesian statistician would assign a probability of 0.5 to the event of heads. Assume now that once a prediction has been made, you are informed that the coin has either two heads or two tails. This is a point of departure between the frequentist and the Bayesian. While the Bayesian, having no other knowledge, is likely to assign a probability of 0.5 to the hypothesis that the coin has two heads, to the frequentist it would seem that no such probability can be assigned. In any case, it would be hard to see why either the Bayesian or the frequentist would have any reason to change their predictions.

Bayes's Theorem, Priors, and Posteriors

The name "Bayesian" comes from the frequent application of Bayes's theorem in Bayesian inference. Bayes's theorem simply relates the marginal and conditional probabilities of two events:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]},$$

when $\Pr[A] > 0$ and $\Pr[B] > 0$. The theorem itself is in no way controversial and is valid under both the frequentist and Bayesian interpretation of probability. The controversy arises with the use of priors and computation of posterior probabilities. Assume that we have a data set D which we believe to be generated by one of n different fully specified models $M_1, \ldots, M_n$, and that the conditional probabilities $\Pr[D|M_i]$ are well-defined and easily computable. To choose among the models, we could for example use maximum likelihood and pick the model for which $\Pr[D|M_i]$ is greatest. The Bayesian approach is instead to compute posterior probabilities on the set of models using Bayes's theorem:

$$\Pr[M_i|D] = \frac{\Pr[D|M_i]\Pr[M_i]}{\Pr[D]}.$$

Using the law of total probability, the denominator of the right hand side of the above equation can be written as

$$\Pr[D] = \sum_{i=1}^{n} \Pr[D, M_i] = \sum_{i=1}^{n} \Pr[D|M_i]\Pr[M_i].$$

The probability $\Pr[M_i]$ is called a prior and represents our belief that model M_i is the correct model *before* observing the data. The probability $\Pr[M_i|D]$ is called the posterior and represents our updated belief that M_i is the correct model *after* having observed the data.

An example of a continuous case is when we want to estimate a parameter of a model. Assume that we know that a distribution M has generated the data, but the parameter θ of M is unknown. The distribution $p[\theta]$ represents how likely different values of θ are prior to having seen the data. After data has been gathered, the posterior distribution of θ is given by Bayes's theorem:

$$p[\theta|D] = \frac{\Pr[D|\theta]p[\theta]}{\int \Pr[D|\theta]\Pr[\theta]d\theta}.$$

We can then estimate θ using the posterior mean $E[\theta|D] = \int \theta p[\theta|D]d\theta$, or we can find an interval (a, b) such that, for given α, $\Pr[a < \theta < b|D] = \int_a^b p[\theta|D]d\theta = 1-\alpha$.

The use of priors continues to be controversial, and a thorough discussion lies outside the scope of this thesis. One objection to the use of priors is that they are not objective, but see [14] for a perspective on objectivity in Bayesian and classic statistics. Another is that it may be difficult to express one's beliefs in terms of a probability distribution and instead one might choose a certain prior for practical reasons rather than to express one's true opinions. And besides, humans often harbor inconsistent opinions that contradict the axioms of probability theory.

The good news is however that priors do not always have a strong influence on the results. It can be shown that under certain conditions, an increasing amount of data will lead to more and more similar posterior distributions [16]. Others have argued for the use of non-informative priors as being an objective choice when there is a lack of prior opinion, see for example [81].

There are often other benefits to Bayesian analysis compared to classical statistics that compensate for the use of priors. For example, in phylogenetics, MCMC techniques have enabled the use of arbitrary prior distributions, and a more efficient investigation of the state space. Bayesian methods allow us to efficiently deal with high dimensional models, and to obtain marginal distributions on the parameters of interest. Also, Bayesian methods allow us to explore posterior distributions rather than just summary statistics such as mean and variance. For an introduction to Bayesian statistics, see the excellent book by Peter Lee [95].

MCMC

In Bayesian inference, we are often confronted with various integration and optimization problems. We frequently need to find a marginal distribution or compute expectations. When dealing with large dimensional spaces, analytic solutions are usually not readily available. Instead, we can obtain Monte Carlo estimates by drawing iid samples from a target distribution. These samples can then be used to approximate the density or expectation of interest. Assume for example that we draw a set of iid samples $x_1, \ldots, x_N$ from a density $p[x]$ defined on a high-dimensional space X. The integral

$$\int_X f(x)p[x]dx$$

can then be approximated by the sum

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

Markov Chain Monte Carlo is a method based on Markov chains that allows us to obtain samples from non-standard distributions from which we cannot draw samples directly. To use MCMC techniques we need to be able to evaluate at least ratios of the target distribution.

First, we describe MCMC on finite state spaces. Assume that the state space is $S = \{s_1, \ldots, s_n\}$. A Markov chain on S is a sequence of random variables $X_1, X_2, \ldots$ taking values in the state space S such that

$$\Pr[X_n | X_{n-1}, \ldots, X_1] = \Pr[X_n | X_{n-1}].$$

The chain is called *homogeneous* if

$$\Pr[X_n = s | X_{n-1} = s'] = \Pr[X_j = s | X_{j-1} = s'],$$

for all $j, n > 1$. In other words, a homogeneous Markov chain lacks memory and the distribution on the states for the next step is fully determined by the current state. A Markov chain on a finite state space can be represented by a transition matrix T, where $T_{ij} = \Pr[X_n = s_j | X_{n-1} = s_i]$ and $\sum_j T_{ij} = 1$. Now, assume that we pick a start state randomly from a distribution represented by a row vector $\pi = (\Pr[s_1], \ldots, \Pr[s_n])$. The distribution on the states after one transition is given by πT. In general, the probability distribution on the states after n transitions is given by πT^n. If the transition matrix is both irreducible and aperiodic, then the distribution on the states will converge to a unique invariant distribution μ irrespective of the start distribution π, i.e.,

$$\pi T^n \to \mu \qquad \text{as } n \to \infty,$$

for any distribution π. A transition matrix is irreducible if any state is reachable from any other in a finite number of steps. A state i has period k if the length of any path returning to i is always a multiple of k. A transition matrix with a state of period $k > 1$ is called periodic. Otherwise, it is called aperiodic.

A sufficient condition to ensure that μ is the desired distribution p from which we want to sample is the so-called *detailed balance* criterion:

$$p_i T_{ij} = p_j T_{ji}.$$

When the state space is infinite, we can instead use the Metropolis-Hastings algorithm [111, 73]. Assume that we want to sample from a distribution on a multi-dimensional state space $\Theta = (\Theta_1, \ldots, \Theta_m)$. The components of Θ can be either discrete, continuous, or a mix of discrete and continuous variables. More generally, each component can be a multi-dimensional random variable in itself. As an example, consider phylogeny where Θ could be a phylogenetic tree together with lengths associated with the edges.

Instead of a transition matrix, the Markov chain is now specified as a set of proposal distributions, $R_k(\Theta'_k|\Theta)$, $k = 1, \ldots, m$, that given the current state Θ propose a change in one of the components Θ_k according to some distribution. In other words, the proposed state Θ' obtained from R_k differs from Θ only in the kth component: $\Theta_i = \Theta'_i$, $i \neq k$. The proposed new state is then accepted with probability $A(\Theta, \Theta')$, in which case Θ' becomes the new current state, or is rejected and the current state remains Θ. The component to change in each step is often picked randomly in each step. In the original Metropolis algorithm, the acceptance probability is

$$A(\Theta, \Theta') = \min\left(1, \frac{\Pr[\Theta']}{\Pr[\Theta]}\right),$$

so that if the probability of the proposed state is greater than the current state, it is always accepted, and otherwise, it is accepted with probability $\Pr[\Theta'] / \Pr[\Theta]$. One way to achieve detailed balance under this acceptance strategy is to ensure that

$$R_k(\Theta'_k|\Theta) = R_k(\Theta_k|\Theta').$$

Hastings generalized the Metropolis algorithm to allow non-symmetric proposal distributions. The acceptance probability is then

$$A(\Theta, \Theta') = \min\left(1, \frac{\Pr[\Theta'] R_k(\Theta_k|\Theta')}{\Pr[\Theta] R_k(\Theta'_k|\Theta)}\right).$$

This again ensures detailed balance.

For more on Metropolis algorithms and MCMC, and discussions on convergence rates, tests of convergence, and more, see for example [120, 98].

Chapter 4

Computational Methods and Models for Duplications and LGTs

This chapter gives an overview of the different phylogenetic methods concerned with gene duplications and LGTs. Algorithms have been developed for a variety of problems, such as tree reconciliation, species tree reconstruction, and orthology analysis. Tree reconciliation is the problem of explaining the differences between a species tree and a corresponding gene tree by giving a plausible evolutionary history of the latter inside the former. Species tree reconstruction refers to problems in which an optimal species tree is sought when given a set of incongruent gene trees. Alternatively, the input can consist of sequences from different gene families, in which case substitution models are also taken into account when seeking optimal species trees. In orthology analysis, the problem is determining whether or not pairs of sequences are orthologous. Data in this case can consist of either trees, sequences, or a mix of trees and sequences.

The first section of this chapter deals with the observation that certain problems in parasitology and biogeography are analogous to those of molecular evolution. The subsequent three sections give a background on previous work on the problems mentioned in the previous paragraph. A common feature of all the methods discussed is their focus on duplications and LGTs. Sections 4.5 and 4.6 discuss the work presented in Papers I, II, and III of this thesis.

4.1 Trees Within Trees

The notion of a tree structure evolving inside another tree structure has been used in at least three separate disciplines: molecular evolution, parasitology, and biogeography. In each case, questions arise about how a *host* is tracked by an *associate*. In molecular evolution, genes track organisms; in parasitology, parasites track hosts; and in biogeography, organisms track areas. The structure most widely used to depict histories of hosts and associates is that of a tree. Different historic events cause the trees of hosts and their associates to be incongruent, thus creating the need to specify exactly what those events are and where they have occurred. As it turns out, each event considered in one discipline has an analogue in the others and results in the same type of incongruity between host and associate trees. The fundamental similarity between the problems in the different disciplines was not recognized until the 1990s, although similar work in the different disciplines had been done independently. A clear exposition on this subject can be found in [129].

In molecular evolution, the events under consideration are speciations, duplications, LGTs, and losses. The corresponding events in parasitology are co-speciation, independent parasite speciation, host switching, and lineage sorting, respectively. In biogeography we have vicariance, sympatry, dispersal, and extinction. The fundamental observation here is that a single model in which an associate tree evolves inside a corresponding host tree is adequate to capture all three cases. In fact, we can find more examples where such a model can be applied; for example, the evolution of protein domains inside gene trees.

In the following sections, we will formulate problems and discuss methods using terms from molecular evolution.

4.2 The Duplication-Loss Model

In the duplication-loss model, we assume that any incongruities between an organismal tree and a corresponding gene tree are due to duplications and losses. In other words, we assume that the mode of genetic transfer is strictly vertical from parent to child, although genes can be duplicated or lost and this change in genetic make-up is sometimes spread to the entire population and is fixed. Clearly, the history of a set of homologous genes represented by a gene tree is then restricted to having occurred inside the edges of a corresponding species tree. Each internal gene tree vertex corresponds to either a duplication event or a speciation event. See Figure 4.1 where a gene tree is drawn inside a species tree showing the evolutionary history of a set of genes; this is an example of tree reconciliation in which a biologically feasible explanation is provided for the disparity between the host and associate trees.

In 1970, Fitch [57] made a distinction between paralogous and orthologous genes, i.e., genes whose least common ancestor in the gene tree is a duplication or speciation, respectively. Similar concepts had been developed much earlier in parasitology, see for example [28]. The development of methods for detecting pairs or groups of orthologous genes is an important step in the prediction of gene function. Traditionally, trees are taken as the data to be analyzed. A species tree is assumed given, together with a gene tree that has been constructed from sequences using methods analogous to those discussed in Section 2.4. Given a species tree and a corresponding gene tree, an important problem is determining the evolutionary history of the gene tree within the species tree and to answer questions such as *which pairs of genes are orthologous and which paralogous?*.

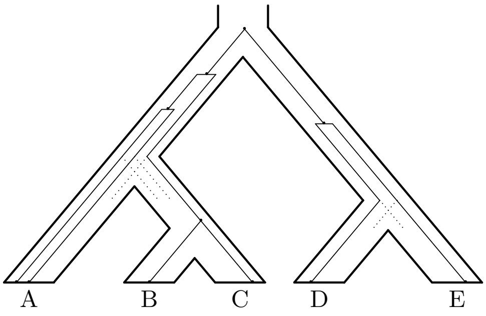

Figure 4.1: ***Example of genes evolving inside a species tree according to the duplication-loss model.***

We will use the following notation in the discussions below. We take all trees to be rooted binary trees. The edges of a rooted tree are assumed to be directed away from the root. The subtree of a tree T rooted at the vertex u is denoted T_u. The correspondence between a species tree S and a gene tree G is given via a leaf-mapping function $\sigma : L(G) \to L(S)$ that maps each gene to the extant species to which it belongs. For convenience, we assume that σ is extended to map sets of gene tree leaves to the corresponding sets of species tree leaves. The function mapping a set of tree vertices to their least common ancestor is denoted lca .

In 1979, Goodman *et al.* gave a parsimony method for tree reconciliation that maps the gene tree inside the species tree such that the number of inferred duplications and losses is minimized. This mapping is called the least common ancestor mapping, $\lambda : V(G) \to V(S)$, and is defined by

$$\lambda(u) = \operatorname{lca}(\sigma(L(G_u))).$$

Note that λ maps vertices of G to vertices of S. When describing a possible evolutionary history G inside S, the gene tree vertices representing speciation events are associated with the species tree vertex that corresponds to the same speciation event. A duplication vertex in G is associated with the edge of S in which the duplication occurred. The interpretation of the mapping given by λ is that if $u \in V(G)$ represents a speciation, then $\lambda(u)$ represents that same speciation event, and if u is a duplication, then the duplication occurred along the incoming edge of $\lambda(u)$.

Building on the framework provided by Goodman *et al.*, Guigó *et al.* attempted to find the species tree whose reconciliation with a set of gene trees requires a minimum number of duplications [65]. Their method can be described as a heuristic local search method or hill-climbing where the neighborhood of a tree is defined by nearest neighbor interchange (NNI) operations [116]. Ma *et al.* proved hardness results for several species tree reconstruction methods [102], and several heuristic methods for species tree reconstruction have been developed [136, 128, 151].

Assuming that at least one gene tree has had a constant number of lineages in each species tree lineage, there is an FPT algorithm for reconstructing the optimal species tree [68].

Going beyond parsimony methods, probabilistic models of gene evolution for the duplication-loss model have recently been proposed. In [4, 5, 139, 155], a complete framework for computational analysis in a probabilistic setting has been developed. The model is most conveniently described as a generative model that generates a gene tree and sequences on a give species tree with times associated with the species tree edges. The model of evolution is based on the standard birth-death process [86] which generates duplications and losses along the edges of a species tree resulting in a gene tree. Sequences are then generated according to an arbitrary choice of standard substitution models. Adopting a Bayesian approach and using MCMC techniques, it is possible to compute various posterior probabilities of interest such as the probability of a gene tree given sequences, or the probability of two sequences being orthologous or paralogous. Posterior probability distributions of duplication and loss rates can also be studied. The latest development in this direction extends the model with an iid model of sequence evolution rate variation across gene tree edges [155]. Methods that simultaneously consider sequence evolution and gene tree and species tree reconciliation when identifying duplications have been termed duplication analysis. An *ad hoc* method for duplication analysis that also takes gene order information into account was presented in [149].

4.3 The Transfer-Loss Model

Early attempts at defining evolutionary models taking LGTs into account include the network model [147, 74, 75]. A related approach considers the subtree transfer operation on a tree in which a subtree is moved to a different location. The corresponding optimization problem is to find a minimal set of subtree transfer operations that transform one given tree to another [34, 33, 76]. Nakhleh *et al.* developed a heuristic for phylogenetic network reconstruction given a species tree and a set of gene trees [119].

Analogous to the case of duplications, tree reconciliation problems have been considered in settings where only transfers and losses are take into account. Variations on parsimony problems were defined and considered in [69].

Probabilistic models have also been suggested. In [82], a model was described in which sequences evolve along a network. Huelsenbeck *et al.* developed a Bayesian framework in the context of hosts and parasites for detecting host switches in which the data to be analyzed consists of host and parasite sequences. The model considers only the case where each host is tracked by a single parasite species so that when a host acquires a new parasite, the parasite formerly associated with the host becomes extinct. In [97], a generative model for LGT, without duplications and losses, based on a Poisson process rather than a birth-death process, was presented and used to generate synthetic data. Biological data was also compared with synthetic data in order to infer LGT rates. In [17], Boc and Makarenkov develop a method for detecting LGTs based on distances between the sequences used to infer the species tree and gene tree. A heuristic algorithm for inferring a network using a minimum number of LGTs from a species tree and a corresponding set of gene trees was developed by Nakhleh *et al.* in [119].

Other, non-phylogenetic, methods for detection of LGTs include the use of atypical sequence composition, which can be used to detect recent LGTs, see for example [9].

4.4 The Duplication-Transfer-Loss Model

Based on the idea of reconciled trees, Charleston developed a computer program, Jungles [26], attempting to solve a parsimony variant of tree reconciliation that considers both duplications and transfers. Unfortunately, the presentation lacks mathematical rigor and is plagued by errors in proofs. The time complexity of the method was not analyzed, and in fact, is probably exponential. Further evidence to support this conjecture is found in Paper I of this thesis where it is shown that finding most parsimonious reconciliations that are temporally feasible is NP-hard.

Although probabilistic models of gene evolution for the duplication-transfer-loss model (DTL-model) have been defined, see for example [32], these have not previously been used to infer transfers or to reconcile trees. A probabilistic model based on the birth-death process, as described in Paper III of this thesis, was in fact used to produce synthetic data for analysis in [67]. In [32], a similar model was suggested, but was applied only to gene family sizes. The two models differ in that the model in [32] assumes that the transfer rate is constant and independent of the number of gene lineages currently present.

4.5 DTL-scenarios

In this section, we discuss the combinatorial model and methods for tree reconciliation presented in Papers I and II. The work is a contribution along the lines of the work of Goodman *et al.* but for the much more complicated case when both duplications, transfers, and losses are considered.

We consider as input a species tree S and a gene tree G, both rooted binary trees. The association of genes with species is given by a leaf-mapping function $\sigma : L(G) \to L(S)$. For convenience, we extend σ to map sets of gene tree leaves to the corresponding sets of species.

Given S, G, and σ, our aim in Paper I is to find the most parsimonious reconciliation explaining the evolution of G with respect to S. Adding LGT as an evolutionary event yields a complexity that requires strict formal definitions. In order to achieve both biological and mathematical soundness in our definitions, we introduce the concept of DTL-scenarios whose associated costs are defined as the number of duplications and LGTs. We do not need to consider every biologically

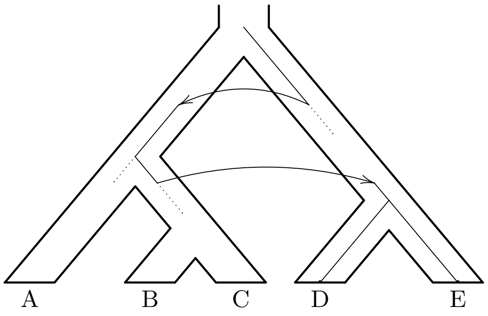

Figure 4.2: ***Example of a non-parsimonious reconciliation in the DTL-model.***

possible reconciliation between S and G in a parsimony setting, and therefore, the definition of DTL-scenarios have been carefully crafted to ensure both biological feasibility as well as non-redundancy. Consider, for example, the evolutionary history shown in Figure 4.2. Although the example is biologically possible, there is no need to consider such histories in a parsimony setting. A thorough justification for our definition is given in Paper II.

A DTL-scenario for S, G, and σ consists of a partition $\{\Sigma, \Delta, \Theta\}$ of the internal vertices of G, a subset $\Xi \subset E(G)$, and a function $\gamma : V(G) \to V(S)$. The subset Ξ consists of all the transfer edges of G. The parts Σ, Δ, and Θ correspond to the speciation, duplication, and transfer vertices of G, respectively. Finally, γ maps the gene tree into the species tree indicating where the speciations, duplications, and lateral gene transfers have occurred. Formally, A DTL-scenario for a species tree S, a gene tree G, and a leaf-mapping function $\sigma : L(G) \to L(S)$ is an octuple

$$(S, G, \sigma, \gamma, \Sigma, \Delta, \Theta, \Xi),$$

where S and G are rooted binary trees, $\sigma : L(G) \to L(S)$ is a leaf-mapping function, $\gamma : V(G) \to V(S)$ is an extension of σ, Σ, Δ, and Θ form a partition of the internal vertices of G, and $\Xi \subset E(G)$ is a subset of the gene tree edges such that:

(I) If u is an internal gene tree vertex with children v and w, then

  a) $\gamma(u)$ is not a proper descendant of $\gamma(v)$ or $\gamma(w)$

  b) At least one of $\gamma(v)$ and $\gamma(w)$ is a descendant of $\gamma(u)$

(II) $(u, v) \in \Xi$ if and only if $\gamma(u)$ is incomparable to $\gamma(v)$

(III) If u is an internal gene tree vertex with children v and w, then

  a) $u \in \Theta$ if and only if $(u, v) \in \Xi$ or $(u, w) \in \Xi$

  b) $u \in \Sigma$ only if $\gamma(u) = \text{lca}\,\{\gamma(v), \gamma(w)\}$ and $\gamma(v)$ and $\gamma(w)$ are incomparable

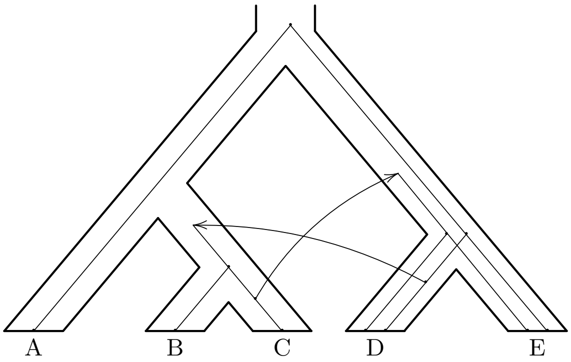

Figure 4.3: ***An example of a cyclic DTL-scenario.***

c) $u \in \Delta$ only if $\gamma(u)$ is an ancestor of lca $\{\gamma(v), \gamma(w)\}$

We also need to consider the fact that some sets of LGTs can lead to temporally infeasible reconciliations, an example of which is shown in Figure 4.3. We say that a DTL-scenario is acyclic if and only if

(V) There is a total order $<$ on $V(S)$ such that

a) if $(x, y) \in E(S)$, then $x < y$

b) if $(u, v), (u', v') \in \Xi$ and v is an ancestor of v', then $p(\gamma(u)) < \gamma(v')$

In Paper II, we prove that the condition of acyclicity given above is both sufficient and necessary to ensure temporal feasibility.

A major result in Paper I is that finding most parsimonious acyclic DTL-scenarios is NP-hard. However, earlier results suggest that in most data-sets cyclicity is usually not a problem. Therefore, dropping the requirement of acyclicity we develop a polynomial-time dynamic programming (DP) algorithm as well as an FPT-algorithm for finding most parsimonious DTL-scenarios. Our algorithms are applied to biological data that have been previously analyzed in the literature with respect to LGTs.

In Paper II, we extend our model by allowing arbitrary costs to be associated with duplications and LGTs and give a DP algorithm for finding minimal-cost DTL-scenarios. The algorithm in Paper II constitutes a considerable improvement on the time complexity of the DP algorithm in Paper I.

For any species tree and gene tree pair, there are only a finite number of DTL-scenarios. The algorithms mentioned so far are able to find optimal DTL-scenarios for any given cost scheme. Due to the combinatorial nature of the problem, there are sets of cost schemes with the same set of optimal DTL-scenarios. An interesting computational problem is to partition the space of cost schemes based on the sets of optimal DTL-scenarios. This is analogous to the problem of parametric sequence alignment [66]. In Paper II, we give a polynomial-time algorithm for parametric tree reconciliation. With this algorithm at our disposal, we are able to obtain the set of all DTL-scenarios that are optimal under *any* cost scheme. We then use this method to perform tests on synthetic data, yielding very encouraging results, that show the trade-off between sensitivity and specificity for different cost schemes.

4.6 A Comprehensive Probabilistic Model of Gene Evolution

In Paper III, we develop a probabilistic model of gene evolution with duplications, LGTs, and losses. To our knowledge, this is the first such probabilistic model that has been used for inference of duplications and LGTs.

We assume that a fixed species tree is given with divergence times associated with its vertices. The model is best described as first generating a gene tree with branch lengths after which some standard substitution model can be used to generate sequences. The model uses a standard birth-death process to generate a gene tree with respect to the species tree given rates for duplications, LGTs, and losses. The resulting gene tree has times associated with its edges. We achieve a relaxed molecular clock by assuming that substitution rates on gene tree edges are iid Γ-distributed variables. The rates obtained from the Γ-distribution, together with edge times, induce branch lengths on the edges of the gene tree. Finally, a substitution model is used to generate sequences. In Paper III, we use the JTT model, but any standard substitution model can be used.

Many interesting computational problems can be defined based on the model described above. We provide a Bayesian framework for the analysis of sequence data using MCMC techniques. We use priors on the parameters of the model, namely the rates of duplication, LGT, and gene loss and the mean and variance of the Γ distribution. A state in our Markov chain is a triple (G, l, θ), where G is a gene tree, l is a function assigning branch lengths to the edges of G, and θ is the set of birth-death rates and the mean and variance of the Γ distribution. By using standard MCMC techniques, interesting posterior distributions can be studied. Examples include the posterior distribution on the gene tree topologies, the LGT or duplication rate, and the number of LGTs.

In order to use MCMC, we need to be able to compute ratios of posterior probabilities of the form $\Pr[G, l, D|\theta]$, where D is the data to be analyzed in the form of sequences (since the species tree is fixed, we omit it from our notation). We can rewrite the probability of a state in our Markov chain as

$$\Pr[G, l, \theta|D] = \frac{\Pr[D|G, l] \Pr[G, l|\theta] \Pr[\theta]}{\Pr[D]}.$$

When computing ratios of posterior probabilities, the denominator in the above expression will cancel, and therefore, we do not need to compute $\Pr[D]$. $\Pr[\theta]$ is simply our prior distribution on the parameters, and $\Pr[D|G, l]$ can be computed according to our chosen substitution model. A major contribution of Paper III is an algorithm for computing the probability $\Pr[G, l|\theta]$. More specifically, we approximate $\Pr[G, l|\theta]$ by introducing discretization points on the species tree and applying a mix of dynamic programming algorithms and techniques from numerical analysis.

Chapter 5

Modeling Cancer Progression

This chapter provides a short background on cancer progression models, and Section 5.2 contains a description of the work presented in Paper IV.

5.1 Overview of Current Methods

Mathematical modeling of cancer progression started more than fifty years ago with simple, yet groundbreaking, models of tumorigenesis [122, 3, 90]. The early models all assumed that cancer is a stochastic multistep process with small transition rates. A more recent example in that direction is [83]. As noted in Section 2.6, cancer progression is an evolutionary process, and therefore, it is not surprising that methods and models from population genetics have been used extensively, see [113] for a review.

In this thesis, we will follow a different line of research, which started with the introduction of Oncogenetic Trees (OTs) by Desper *et al.* [38]. Since Vogelstein's path model of colon cancer, numerous narrative models for progression of diverse cancer types have been suggested, for example [80, 141, 146]. Such models are often the result of *ad hoc* handmade reconstructions. The introduction of OTs was an attempt at a more stringent mathematical modeling of cancer progression. An OT is a rooted tree where each vertex represents a specific genetic aberration and there is a probability associated with each edge. An OT generates a set of aberrations by first choosing a set of edges, each independently and according to its associated probability. The set of vertices reachable from the root, using only the chosen edges, is then the set of generated aberrations. In this way, an OT induces a probability distribution on the power set of all aberrations.

Given cross-sectional data, i.e., sets of aberrations where each set is from a unique tumor or patient, the computational task is to find the correct OT. Desper *et al.* showed that computing a specific weight function on the set of all pairs of vertices and then using Edmonds's maximum branching algorithm to obtain the topology, the correct tree will be recovered with high probability.

One problem in OTs is that once progression stops at some vertex u, i.e., when none of the outgoing edges of u are chosen in the first step, then progression cannot reach any of the descendants of u. Biological data is almost always noisy, and in any case, real cancer progression is not tree like. The result is that usually every OT, except the OTs with a star topology, assign zero probability to some of the data, and therefore, using likelihood methods is not straightforward. Another problem is that cancer progression is best described using acyclic graphs that allow an aberration to be obtained via different pathways.

In an attempt to capture more of the graph-like progression of cancers, Beerenwinkel *et al.* used mixtures of oncogenetic trees [12, 13, 130]. In order to assign positive probabilities to all data points in the input, the topology of the first OT was kept a star tree. For inference of mixtures, they developed an EM-like algorithm, which has not been proven to deliver locally optimal ML solutions.

5.2 Hidden-variable Oncogenetic Trees

In Paper IV, we introduce Hidden-variable Oncogenetic Trees (HOTs) and mixtures thereof (HOT-mixtures) in an attempt to remedy some of the problems with traditional OTs, while taking advantage of the simplicity of tree structures.

A HOT is a tree where each vertex is associated with a pair of hidden and visible variables. The hidden variable indicates true progression while the visible variable indicates the outcome of a specific experiment, e.g., the absence or presence of a genetic aberration. The hidden and visible variables associated with a vertex u are denoted $Z(u)$ and $X(u)$, respectively. The values of all variables are assumed to be zero (absence) or one (presence). The distribution on the values of each variable is determined by two conditional probability distributions so that a total of four conditional distributions are associated with each vertex:

$$\begin{aligned}
&\Pr[X(u)|Z(u)=0],\\
&\Pr[X(u)|Z(u)=1],\\
&\Pr[Z(u)|Z(p(u))=0],\\
&\Pr[Z(u)|Z(p(u))=1],
\end{aligned}$$

where $p(u)$ denotes the parent of u. Note that the visible variable at a vertex depends only on the hidden variable at the same vertex, and that the hidden variable only depends on the hidden variable of the parent.

When generating a set of aberrations, the values of the hidden variables are determined first. This is similar to oncogenetic trees, except that the hidden variable of a vertex can receive the value one even if its parent has not. The probability $\Pr[Z(u)=1|Z(p(u))=0]$, which is normally small, can be interpreted as the probability that an event associated with a later stage of progression occurs spontaneously although the stages that directly precede it have not been reached. Once the values of the hidden variables are determined, the visible variables receive their values. The probability $\Pr[X(u) = 1|Z(u) = 0]$ is interpreted as the probability of a false positive and the probability $\Pr[X(u) = 0|Z(u) = 1]$ as a false negative. The latter can also include the probability that the progression has reached u but via a different set of events than the aberration associated with u. The set of aberrations generated are the aberrations associated with the vertices whose visible variables have value one.

Paper IV also includes a description of HOT-mixtures. A HOT-mixture consists of a set of HOTs, $\mathcal{T}_1, \ldots, \mathcal{T}_n$, together with a probability distribution on the same. To generate data from a mixture, we first chose a HOT according to the given probability distribution and then generate a set of aberrations from the chosen HOT.

Global structural EM algorithms for inferring HOTs and HOT-mixtures constitute the major computational contributions of Paper IV.

Chapter 6

Overview of Included Articles and Manuscripts

Paper I: We define a combinatorial model for the reconciliation of gene and species trees using gene duplication, lateral gene transfer, and gene loss. A reconciliation is said to be cyclic if its set of transfers are temporally infeasible. We prove that finding most parsimonious acyclic reconciliations is NP-hard. However, simulations have previously shown that in most cases the most parsimonious reconciliations are acyclic. Dropping the requirement of acyclicity, we provide efficient algorithms for construction of most parsimonious reconciliations. We also analyze a biological dataset with our tools and show that our methods work well in practice.

Paper II: We continue to build on the framework provided by our model in Paper I. A thorough discussion on the soundness of our model presented in Paper I is provided. Next, we extend our model to allow arbitrary costs to be associated with duplications and lateral gene transfers and develop efficient methods for finding minimal-cost reconciliations. Analogous to parametric sequence alignment, we derive polynomial-time algorithms for parametric tree reconciliation. Tests are performed on synthetic data that show the performance of our methods.

Paper III: Going beyond combinatorial methods, we define a comprehensive probabilistic model of gene evolution that incorporates a birth-death process generating duplications, lateral gene transfers, and losses, together with a substitution model with a relaxed molecular clock. To our knowledge, this is the first probabilistic model used to simultaneously infer duplications and lateral gene transfers, and is more advanced than any probabilistic model that includes LGT as an evolutionary event. We present methods based on MCMC, numerical analysis, and dynamic programming for computing various posterior distributions and probabilities, including the distribution of rates, gene tree topologies, and counts of lateral gene transfer events.

Paper IV: We define Hidden-variable Oncogenetic Trees (HOTs) and mixtures thereof (HOT mixtures) to capture cancer progression pathways. Vertices of a HOT represent specific genetic aberrations and a pair of hidden and visible variables are associated with each vertex. The hidden variables indicate true progression, while the visible variables indicate outcome of experiments for detection of specific aberrations. Global structural EM algorithms are presented for maximum likelihood estimation of HOTs and HOT mixtures from cross sectional data. Analysis of the performance of our methods on synthetic as well as biological data are presented.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 2007.

[2] D.J. Araten, D.W. Golde, R.H. Zhang, H.T. Thaler, L. Gargiulo, R. Notaro, and L. Luzzatto. A quantitative measurement of the human somatic mutation rate. *Canc res*, 65(18):8111, 2005.

[3] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*, 8(1):1–12, Mar 1954.

[4] L. Arvestad, A.C. Berglund, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Proceedings of the eighth annual international conference on research in computational molecular biology*, pages 326–335, 2004.

[5] L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its associated probabilities. *J ACM*, 56(2):1–44, 2009.

[6] K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2):251–278, 1999.

[7] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and approximation.* Springer New York, 1999.

[8] O.T. Avery, C.M. MacLeod, and M. McCarty. Chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med*, 79:137–58, 1944.

[9] R.K. Azad and J.G. Lawrence. Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res*, 35(14): 4629–39, 2007.

[10] J.A. Bailey and E.E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–564, 2006.

[11] E. Bapteste, E. Susko, J. Leigh, D. MacLeod, R.L. Charlebois, and W.F. Doolittle. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol*, 5(1):33, 2005.

[12] N. Beerenwinkel, J. Rahnenfuhrer, M. Daumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*, 12(6):584–598, Jul 2005.

[13] N. Beerenwinkel, J. Rahnenfuhrer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, May 2005.

[14] J.O. Berger and D.A. Berry. Statistical analysis and the illusion of objectivity. *Am Sci*, 76(2):159–165, 1988.

[15] M.J. Bissell and D. Radisky. Putting tumours in context. *Nat Rev Genet*, 1 (1):46–54, 2001.

[16] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Ann Math Stat*, pages 882–886, 1962.

[17] A. Boc and V. Makarenkov. New efficient algorithm for detection of horizontal gene transfer events. In *Algorithms in Bioinformatics: Third International Workshop, WABI 2003, Budapest, Hungary, September 15-20, 2003: Proceedings*, page 190. Springer, 2003.

[18] C.B. Bridges. Duplication. *Anat Rec*, 15:357–358, 1918.

[19] C.B. Bridges. Salivary chromosome maps. *J Hered*, 26:60–64, 1935.

[20] J.R. Brown. Ancient horizontal gene transfer. *Nat Rev Genet*, 4(2):121–132, 2003.

[21] T.A. Brown. *Genomes.* John Wiley and Sons, Inc., 2002.

[22] F.G. Brunet, H.R. Crollius, M. Paris, J.M. Aury, P. Gibert, O. Jaillon, V. Laudet, and M. Robinson-Rechavi. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol*, 23(9):1808–1816, 2006.

[23] A. Caignard, M.S. Martin, M.F. Michel, and F. Martin. Interaction between two cellular subpopulations of a rat colonic carcinoma when inoculated to the syngeneic host. *Int J Canc*, 36(2), 1985.

[24] M.A.A. Castro, T.T.G. Onsten, R.M.C. de Almeida, and J.C.F. Moreira. Profiling cytogenetic diversity with entropy-based karyotypic analysis. *J Theor Biol*, 234(4):487–495, 2005.

[25] L.L. Cavalli-Sforza and A.W. Edwards. Phylogenetic analysis. models and estimation procedures. *Am J Hum Genet*, 19(3 Pt 1):233–257, May 1967.

[26] M.A. Charleston. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math Biosci*, 149(2):191, 1998.

[27] J. Cheetham, F. Dehne, A. Rau-Chaplin, U. Stege, and P.J. Taillon. Solving large FPT problems on coarse-grained parallel machines. *J Comput Syst Sci*, 67(4):691–706, 2003.

[28] T. Clay. Some problems in the evolution of a group of ectoparasites. *Evolution*, pages 279–299, 1949.

[29] P.F. Cliften, R.S. Fulton, R.K. Wilson, and M. Johnston. After the duplication: gene loss and adaptation in Saccharomyces genomes. *Genetics*, 172(2): 863–872, 2006.

[30] B. Crespi and K. Summers. Evolutionary biology of cancer. *Trends Ecol Evol*, 20(10):545–552, 2005.

[31] F.H. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[32] M. Csűrös and I. Miklós. A probabilistic model for gene content evolution with duplication, loss and horizontal transfer. In *In Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 206–220. Springer, 2006.

[33] B. DasGupta, X. He, T. Jiang, M. Li, and J. Tromp. On the linear-cost subtree-transfer distance between phylogenetic trees. *Algorithmica*, 25(2): 176–195, 1999.

[34] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On distances between phylogenetic trees. In *Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, pages 427–436. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1997.

[35] W.H. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bull Math Biol*, 49(4):461–467, 1987.

[36] P. Dehal and JL Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10):e314, 2005.

[37] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Ser B*, pages 1–38, 1977.

[38] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schaffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*, 6(1):37–51, Spr 1999.

[39] F.X. Diebold and C. Li. Forecasting the term structure of government bond yields. *J Econometrics*, 130(2):337–364, 2006.

[40] D.S. Dolberg, R. Hollingsworth, M. Hertle, and M.J. Bissell. Wounding and its role in RSV-mediated tumor formation. *Science*, 230(4726):676, 1985.

[41] W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–9, 1999.

[42] W.F. Doolittle and E. Bapteste. Pattern pluralism and the tree of life hypothesis. *Proc Natl Acad Sci U S A*, 104(7):2043, 2007.

[43] R.G. Downey and M.R. Fellows. *Parameterized complexity*. Springer Verlag, 1999.

[44] R.C. Edgar and S. Batzoglou. Multiple sequence alignment. *Curr Opin Struct Biol*, 16(3):368–373, Jun 2006.

[45] A.W.E. Edwards and L.L. Cavalli-Sforza. The reconstruction of evolution. *Ann Hum Genet*, 27:105–106, 1963.

[46] W. Edwards, H. Lindman, and L.J. Savage. Bayesian statistical inference for psychological research. *Psychol Rev*, 70(3):193–242, 1963.

[47] I. Elias and J. Lagergren. Fast neighbor joining. *Theor Comput Sci*, 410 (21–23):1993–2000, 2009.

[48] J.S. Farris. Inferring phylogenetic trees from chromosome inversion data. *Syst Zool*, 27:275–284, 1978.

[49] M.R. Fellows. On the complexity of vertex set problems. Technical report, Technical report, Computer Science Department, University of New Mexico, 1988.

[50] M.R. Fellows and M.A. Langston. Nonconstructive advances in polynomial-time complexity. *Inf Process Lett*, 26(3):155–162, 1987.

[51] J. Felsenstein. Alternative methods of phylogenetic inference and their interrelationship. *Syst Zool*, 28:49–62, 1979.

[52] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[53] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates Sunderland, 2003.

[54] J. Felsenstein and G.A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*, 13(1):93–104, Jan 1996.

[55] J.A. Fessler and A.O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Trans Signal Process*, 42(10):2664–2677, 1994.

[56] Crick F.H. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138. Symp Soc Exp Biol, 1958.

[57] W.M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, pages 99–113, 1970.

[58] W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–284, Jan 1967.

[59] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[60] N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *proc 14th Int Conf on Machine Learning*, page 125. Morgan Kaufmann Pub, 1997.

[61] N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural em algorithm for phylogenetic inference. *J Comp Biol*, 9(2):331–353, 2002.

[62] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Freeman San Francisco, 1979.

[63] O. Gascuel. Concerning the NJ algorithm and its unweighted version, UNJ. In B. Mirkin, F. McMorris, F. Roberts, and A. Rhetsky, editors, *Mathematical Hierarchies and Biology*, pages 149–170. AMS, Providence, 1997.

[64] J.P. Gogarten, W.F. Doolittle, and J.G. Lawrence. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, 19(12):2226–2238, 2002.

[65] R. Guigó, I. Muchnik, and T.F. Smith. Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol*, 6(2):189–213, 1996.

[66] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[67] M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 347–356. ACM New York, NY, USA, 2004.

[68] M.T. Hallett and J. Lagergren. New algorithms for the duplication-loss model. *Proceedings of the fourth annual international conference on computational molecular biology*, pages 138–146, 2000.

[69] M.T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. *Proceedings of the fifth annual international conference on computational biology*, 2001.

[70] A.T. Hamilton, S. Huntley, M. Tran-Gyamfi, D.M. Baggott, L. Gordon, and L. Stubbs. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res*, 16(5):584–594, 2006.

[71] D. Hanahan and R.A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Feb 2000.

[72] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

[73] W.K. Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[74] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math Biosci*, 98(2):185–200, 1990.

[75] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol*, 36(4):396–405, 1993.

[76] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Appl Math*, 71(1-3):153–169, 1996.

[77] A.D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol*, 36(1):39–56, 1952.

[78] J.P. Huelsenbeck, F. Ronquist, R. Nielsen, and J.P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550): 2310–2314, Dec 2001.

[79] M. Jamshidian and R.I. Jennrich. Acceleration of the EM algorithm by using quasi-Newton methods. *J Roy Stat Soc B*, pages 569–587, 1997.

[80] J.A. Jankowski, N.A. Wright, S.J. Meltzer, G. Triadafilopoulos, K. Geboes, A.G. Casson, D. Kerr, and L.S. Young. Molecular evolution of the metaplasia-dysplasia-adenocarcinoma sequence in the esophagus. *Am J Pathol*, 154(4): 965–973, Apr 1999.

[81] E.T. Jaynes. Prior probabilities. *IEEE Trans Syst Sci Cybern*, 227, 1968.

[82] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21):2604, 2006.

[83] S. Jones, W.D. Chen, G. Parmigiani, F. Diehl, N. Beerenwinkel, T. Antal, A. Traulsen, M.A. Nowak, C. Siegel, V.E. Velculescu, K.W. Kinzler, B. Vogelstein, J. Willis, and S.D. Markowitz. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci U S A*, 105(11):4283–4288, Mar 2008.

[84] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In M.N. Munro, editor, *Mammalian protein metabolism*, volume 3, pages 21–132. New York, 1969.

[85] P.J. Keeling and J.D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, 9(8):605–618, 2008.

[86] David G. Kendall. On the generalized “birth-and-death” process. *Ann Math Stat*, 19:1–15, 1948.

[87] K.K. Kidd and L.A. Sgaramella-Zonta. Phylogenetic analysis: concepts and methods. *Am J Hum Genet*, 23(3):235–252, May 1971.

[88] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.

[89] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*, 29(2):170–179, Aug 1989.

[90] A.G. Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–823, Apr 1971.

[91] S. Kumar and A. Filipski. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res*, 17(2):127–135, Feb 2007.

[92] J.G. Lawrence. Horizontal and vertical gene transfer: The life history of pathogens. *Contrib Microbiol*, 12:255–271, 2005.

[93] J. Lederberg and E. Tatum. Gene recombination in Escherichia coli. *Nature*, 158:558, October 1946.

[94] J. Lederberg and E.L. Tatum. Novel genotypes in mixed cultures of biochemical mutants of bacteria. In *Cold Spring Harbor Symp. Quant. Biol*, volume 11, pages 113–114, 1946.

[95] P.M. Lee. *Bayesian Statistics: An Introduction.* John Wiley, 2004.

[96] C. Lengauer, K.W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[97] S. Linz, A. Radtke, and A. von Haeseler. A likelihood framework to measure horizontal gene transfer. *Mol Biol Evol*, 24(6):1312–1319, Jun 2007.

[98] J.S. Liu. *Monte Carlo strategies in scientific computing.* Springer, 2001.

[99] L.A. Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Canc Res*, 51(12):3075–3079, 1991.

[100] M. Lynch and J.S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[101] M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000.

[102] B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM J Comput*, 30(3):729–752, 2000.

[103] C.C. Maley, P.C. Galipeau, J.C. Finley, V.J. Wongsurawat, X. Li, C.A. Sanchez, T.G. Paulson, P.L. Blount, R.A. Risques, P.S. Rabinovitch, *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet*, 38(4):468–473, 2006.

[104] C.C. Maley, P.C. Galipeau, X. Li, C.A. Sanchez, T.G. Paulson, P.L. Blount, and B.J. Reid. The combination of genetic instability and clonal expansion predicts progression to esophageal adenocarcinoma. *Canc Res*, 64(20):7629–33, 2004.

[105] C.C. Maley, P.C. Galipeau, X. Li, C.A. Sanchez, T.G. Paulson, and B.J. Reid. Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Canc Res*, 64(10):3414, 2004.

[106] T. Marques-Bonet, J.M. Kidd, M. Ventura, T.A. Graves, Z. Cheng, L.D.W. Hillier, Z. Jiang, C. Baker, R. Malfavon-Borja, L.A. Fulton, *et al.* A burst of segmental duplications in the genome of the african great ape ancestor. *Nature*, 457(7231):877–881, 2009.

[107] G. Mendel. Versuche über Pflanzen-Hybriden. *Verb. Naturforsch. Ver. Brunn*, 4:3–47, 1866.

[108] G. Mendel. *Experiments in plant hybridisation.* Cosimo Classics, 2008.

[109] Xiao-Li Meng and David van Dyk. The EM algorithm–an old folk-song sung to a fast new tune. *J Roy Stat Soc Ser B*, 59(3):511–567, 1997.

[110] X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[111] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J Chem Phys*, 21 (6):1087–1091, 1953.

[112] A. Meyer and Y. Van de Peer. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, 27(9):937–945, 2005.

[113] F. Michor, Y. Iwasa, and M.A. Nowak. Dynamics of cancer progression. *Nat Rev Genet*, 4(3):197–205, 2004.

[114] B.E. Miller, F.R. Miller, J. Leith, and G.H. Heppner. Growth interaction in vivo between tumor subpopulations derived from a single mouse mammary tumor. *Canc res*, 40(11):3977, 1980.

[115] S.H. Moolgavkar and E.G. Luebeck. Multistage carcinogenesis and the incidence of human cancer. *Gene Chromosome Canc*, 38(4), 2003.

[116] G.W. Moore, M. Goodman, and J. Barnabas. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *J Theor Biol*, 38(3):423, 1973.

[117] H.J. Muller. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica*, 17(3):237–252, 1935.

[118] H.J. Muller. A viable two-gene deficiency: phenotypically resembling the corresponding hypomorphic mutations. *J Hered*, 26(11):469, 1935.

[119] L. Nakhleh, D. Ruths, and L.S. Want. RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. *Lecture notes in computer science*, pages 84–93, 2005.

[120] R.M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[121] C.R. Nelson and A.F. Siegel. Parsimonious modeling of yield curves. *J Bus*, pages 473–489, 1987.

[122] C.O. Nordling. A new theory on cancer-inducing mechanism. *Br J Cancer*, 7(1):68–72, Mar 1953.

[123] C. Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3(8):e123, Aug 2007.

[124] P.C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194 (4260):23–28, 1976.

[125] K. Ochiai, T. Yamanaka, K. Kimura, and O. Sawada. Inheritance of drug resistance (and its transfer) between Shigella strains and between Shigella and E. coli strains. *Nihon Iji Shimpo*, 1861:34, 1959.

[126] H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[127] S. Ohno. *Evolution by gene duplication.* Allen and Unwin, 1970.

[128] R.D. Page. Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14(9):819–820, 1998.

[129] R.D.M. Page and M.A. Charleston. Trees within trees: phylogeny and historical associations. *Trends Ecol Evol*, 13(9):356–359, 1998.

[130] J. Rahnenfuhrer, N. Beerenwinkel, W.A. Schulz, C. Hartmann, A. von Deimling, B. Wullich, and T. Lengauer. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10): 2438–2446, May 2005.

[131] MJ Renan. How many mutations are required for tumorigenesis? implications from human cancer data. *Mol Carcinog*, 7(3):139, 1993.

[132] F. Ronquist and J.P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, Aug 2003.

[133] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol*, 10(5):1073–1095, Sep 1993.

[134] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[135] R.V. Samonte and E.E. Eichler. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*, 3(1):65–72, 2002.

[136] M. Sanderson and M. McMahon. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol Biol*, 7(Suppl 1):S3, 2007.

[137] D. Sankoff. Minimal mutation trees of sequences. *SIAM J Appl Math*, 28: 35–42, 1975.

[138] D. Sankoff and P. Rousseau. Locating the vertices of a steiner tree in an arbitrary metric space. *Math Program*, 9:240–246, 1975.

[139] B. Sennblad and J. Lagergren. Probabilistic orthology analysis. *submitted*, 2008.

[140] M.H. Sieweke and M.J. Bissell. The tumor-promoting effect of wounding: a possible role for TGF-beta-induced stromal alterations. *Crit Rev Oncog*, 5 (2-3):297, 1994.

[141] P.T. Simpson, J.S. Reis-Filho, T. Gale, and S.R. Lakhani. Molecular evolution of breast cancer. *J Pathol*, 205(2):248–254, Jan 2005.

[142] P.H.A. Sneath and R.R. Sokal. *Numerical taxonomy: The principles and practice of numerical classification.* W.H. Freeman, San Fransisco, 1973.

[143] M.N. Swartz. Use of antimicrobial agents and drug resistance, 1997.

[144] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3):512–526, May 1993.

[145] J.S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, 38:615–43, 2004.

[146] A.A. van Tilborg, A. de Vries, M. de Bont, L.E. Groenfeld, T.H. van der Kwast, and E.C. Zwarthoff. Molecular evolution of multiple recurrent cancers of the bladder. *Hum Mol Genet*, 9(20):2973–2980, Dec 2000.

[147] A. von Haeseler and G.A. Churchill. Network models for sequence evolution. *J Mol Evol*, 37(1):77–85, 1993.

[148] T.L. Wang, C. Rago, N. Silliman, J. Ptak, S. Markowitz, J.K.V. Willson, G. Parmigiani, K.W. Kinzler, B. Vogelstein, and V.E. Velculescu. Prevalence of somatic alterations in the colorectal cancer cell genome. *Proc Natl Acad Sci U S A*, 99(5):3076, 2002.

[149] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61, 2007.

[150] J.D. Watson and F.H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[151] A. Wehe, M.S. Bansal, J.G. Burleigh, and O. Eulenstein. Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13):1540–1541, 2008.

[152] D.A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.J. Chen, V. Makhijani, G.T. Roth, *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189): 872–876, 2008.

[153] C.R. Woese. On the evolution of cells. *Proc Natl Acad Sci U S A*, 99(13): 8742–7, 2002.

[154] J. Zhang. Evolution by gene duplication: an update. *Trends Ecol Evol*, 18 (6):292–298, 2003.

[155] Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*, 106(14):5714–5719, 2009.